

國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

透過解耦學習影片問答中的時空間關係

Learning by Decoupling Spatial and Temporal Relations
for Video Question Answering

李信穎

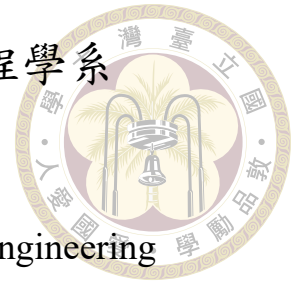
Hsin-Ying Lee

指導教授: 徐宏民 博士

Advisor: Winston H. Hsu Ph.D.

中華民國 111 年 8 月

August, 2022



國立臺灣大學碩士學位論文
口試委員會審定書

透過解耦學習影片問答中的時空間關係

Learning by Decoupling Spatial and Temporal Relations for
Video Question Answering

本論文係李信穎君（學號 R09922084）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 111 年 8 月 12 日承下列考試委員審查通過及口試及格，特此證明

口試委員：



陳文進 (指導教授)

陳奕廷

葉福玗

系主任

洪士瀨





Acknowledgements

感謝徐宏民教授的指導，感謝同組蘇弘庭學長、蔡秉辰和實驗室同學吳宗翰、林承緯、葉佳峯學長陪我討論和給我很多研究上和寫作上的建議。也感謝CMLab實驗室提供豐富的資源。





摘要

雖然最近大規模的影片語言預訓練在影片問答方面有了很大的進展，但空間建模的設計沒有圖像語言模型的那麼精緻；現有的時間建模方式也受到模態之間沒有對齊的影響。為了學習精緻的視覺理解，我們將時空建模解耦，並提出了一種結合圖像和影片語言編碼器的混合結構。前者獨立於時間從較大但稀疏採樣的影格中理解空間語義，而後者在較低空間但較高時間解析度下捕捉時間動態。另外，為了幫助影片語言模型學習影片問答的時間關係，我們提出了一種新穎的預訓練目標，即時間引用建模，它要求模型辨別影片序列中事件的時間位置。透過廣泛且詳細的實驗，我們證明這個方法做得比以前在數量級更大的資料集上預訓練的研究更好。

關鍵字：機器學習、深度學習、影片理解、時空間推理、影片問答





Abstract

While recent large-scale video-language pre-training made great progress in video question answering, the design of spatial modeling is less fine-grained than that of image-language models; existing practices of temporal modeling also suffer from weak and noisy alignment between modalities. To learn fine-grained visual understanding, we decouple spatial-temporal modeling and propose a hybrid pipeline integrating an image- and a video-language encoder. The former encodes spatial semantics from larger but sparsely sampled frames independently of time, while the latter models temporal dynamics at lower spatial but higher temporal resolution. To help the video-language model learn temporal relations for video QA, we propose a novel pre-training objective, Temporal Referring Modeling, which requires the model to identify temporal positions of events in video sequences. Extensive and detailed experiments demonstrate that our model outperforms previous work that pre-trained on orders of magnitude larger datasets.

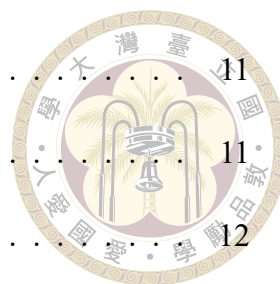
Keywords: Machine Learning, Deep Learning, Video Understanding, Spatial-Temporal Modeling, Video Question Answering





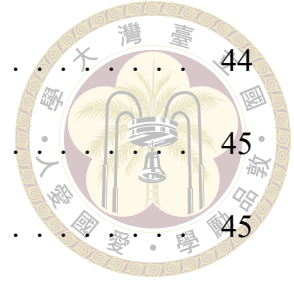
Contents

	Page
Verification Letter from the Oral Examination Committee	ii
Acknowledgements	iii
摘要	v
Abstract	vii
Contents	ix
List of Figures	xiii
List of Tables	xv
Chapter 1 Introduction	1
Chapter 2 Related Work	5
2.1 Video Question Answering	5
2.2 Pre-training for Temporal Relation Modeling	5
2.2.1 Learning from Global Alignment.	6
2.2.2 Learning from Local Alignment and Frame Ordering.	6
2.2.3 Learning from Large-Scale Video Question Answering Datasets. . .	6
2.3 Encoding Motion and Appearance	7
Chapter 3 Method	9
3.1 Decoupled Spatial-Temporal Encoders	9



3.1.1	Image-Language Encoding.	11
3.1.2	Video-Language Encoding.	11
3.1.3	Answer Selection.	12
3.2	Temporal Referring Modeling	12
Chapter 4	Experiments	15
4.1	Preliminary Analysis	15
4.1.1	Encoding Spatial Semantics	15
4.1.2	Modeling Temporal Relationships	16
4.2	Video Question Answering	18
4.3	Ablation Studies	19
Chapter 5	Conclusion	21
References		23
Appendix A — Implementation Details		37
A.1	Model Architectures	37
A.2	Video-Language Pre-training	38
A.2.1	Details of Question and Video Synthesis for Temporal Referring Modeling	38
A.2.2	Auxiliary Objective with Contrastive Learning	39
A.2.3	Pre-training Datasets	40
A.3	Optimization	41
Appendix B — Experiment Details		43
B.1	Details of Temporal Modeling Analysis	43
B.2	Pre-training Data Used by Prior Approaches	44

B.3	Full Results and Analysis on AGQA 2.0	44
B.3.1	Full Results of Temporal Modeling Analysis	45
B.3.2	Full Results and Analysis of Our Method	45
B.3.3	Full Results of Ablation Study of Encoding Streams	48







List of Figures

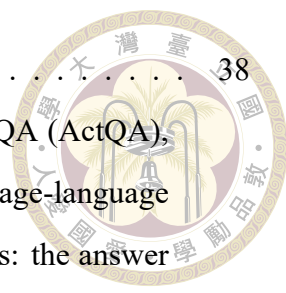
1.1	Comparison between (a) previous and (b)(c) our approaches for video QA. (a) Prior work solved video QA by video-language pre-training but might suffer from lack of event details, video-transcript misalignment or limited diversity of pre-training questions. (b) We pre-train a video-language encoder to learn event representations and temporal relations between them by asking the model to identify specific events in synthesized video sequences. (c) We integrate the video-language model with a pre-trained image-language model to encode fine-grained spatial and temporal semantics at different spatial-temporal resolutions.	2
3.1	Our video QA pipeline. Encoded questions are fused with frames and videos to gather spatial and temporal information. Their representations are then compared with answer candidates to obtain the final predictions. (* marks the frozen modules.)	10
3.2	Temporal Referring Modeling, which associates visual events and their temporal relationships with languages by asking absolute and relative positions of events in concatenated video features sampled from video captioning data.	13





List of Tables

4.1	Comparison between prior methods and our upper bound of ActivityNet-QA by question type. ALBEF exhibits advantages on the questions involving spatial reasoning. (Rel. is short for Relationships, and UB is the abbreviation for upper bound.)	16
4.2	Results of prior work taking shuffled frames as input. The little performance drop indicates that some methods are not sensitive to the order of frames. (* signifies that input frames are shuffled. We report the average of three results for the shuffle experiment.)	17
4.3	Comparison with previous methods on ActivityNet-QA. We outperform all methods with significantly less pre-training data. The dataset names are provided in the supplement. (img: images. vid: videos.)	18
4.4	Comparison with prior methods on ActivityNet-QA by question type. We perform comparably in question types of spatial information and improve temporal modeling.	18
4.5	Comparison with prior work on AGQA 2.0. We list the best performance among methods without (Best w/o PT) and with pre-training (Best w/ PT) for each question type. Ours exceeds all methods in all question types. (Rel.: Relationships. Compar.: Comparison. Recog.: Recognition.)	19
4.6	Ablation study of input modalities and pre-training strategies on AGQA 2.0. The results favor our hybrid pipeline and TRM. (✓ means the modality is presented. VQA: pretrained on VQA. TRM: pre-trained with TRM. *: shuffled input.)	20
4.7	Ablation study of two encoding streams on AGQA 2.0.	20



A.1	Hyperparameters for the architecture.	38
A.2	Hyperparameters for pre-training (Pre-Train), ActivityNet-QA (ActQA), and AGQA 2.0 (AGQA). Base: the question, image, and image-language encoder. Video: the video and video-language encoder. Ans: the answer encoder.	41
B.3	Results of VIOLET taking shuffled frames as input on the questions of State Transition of TGIF-QA. (* signifies that input frames are shuffled. We report the average of three results for the shuffle experiment.)	44
B.4	Reasoning types and examples of their templates of AGQA 2.0.	45
B.5	Full results of the preliminary analysis of temporal modeling on AGQA 2.0. (* means shuffled input. We report the result of one experiment.) . .	46
B.6	Full results of our method on AGQA 2.0 with ablation of components and pre-training strategies. (T: questions; F: frames; F : frames with the image-language encoder pre-trained on VQA; V: videos; V : videos with the video-language encoder pre-trained with TRM; *: shuffled video inputs.) 47	
B.7	Full results of the ablation study on two encoding streams. (IL: image-language encoder; VL: video-language encoder.)	48



Chapter 1 Introduction

Videos are the complex composition of human actions, objects, scenes, and their interactions over time. To examine the capability of machines for video understanding, video question answering (video QA), a task of answering questions about videos, is proposed and requires machines to associate questions in natural languages with visual contents, including scenes [76, 84], dialogues [8, 37], temporal relationships [21, 27, 74, 85], and higher-order cognition [38, 74, 82]. Recent breakthroughs were achieved by pre-training a deep multi-modality encoder, mostly Transformer [70], with large-scale video-language datasets [3, 50, 79]. Models first learned semantic connections between visual and linguistic contents and then were fine-tuned on downstream video-language tasks [41, 59, 79, 86, 91].

Despite the advance of this framework in video QA, the spatial semantics encoding of video-language (VL) models is not as fine-grained as the sophisticated design for image-language (IL) models [2, 60, 88]. A preliminary analysis shows that on video QA benchmarks entailing spatial and temporal knowledge, simply averaging frame-by-frame predictions of an IL model can sometimes outperform state-of-the-art VL models. Though the VL models exhibit a slight advantage in questions involving temporal information, the IL model greatly excels in capturing spatial clues (improvement by 7% accuracy; see the full results in Section 4.1.1). The positive performance of IL models could also be at-

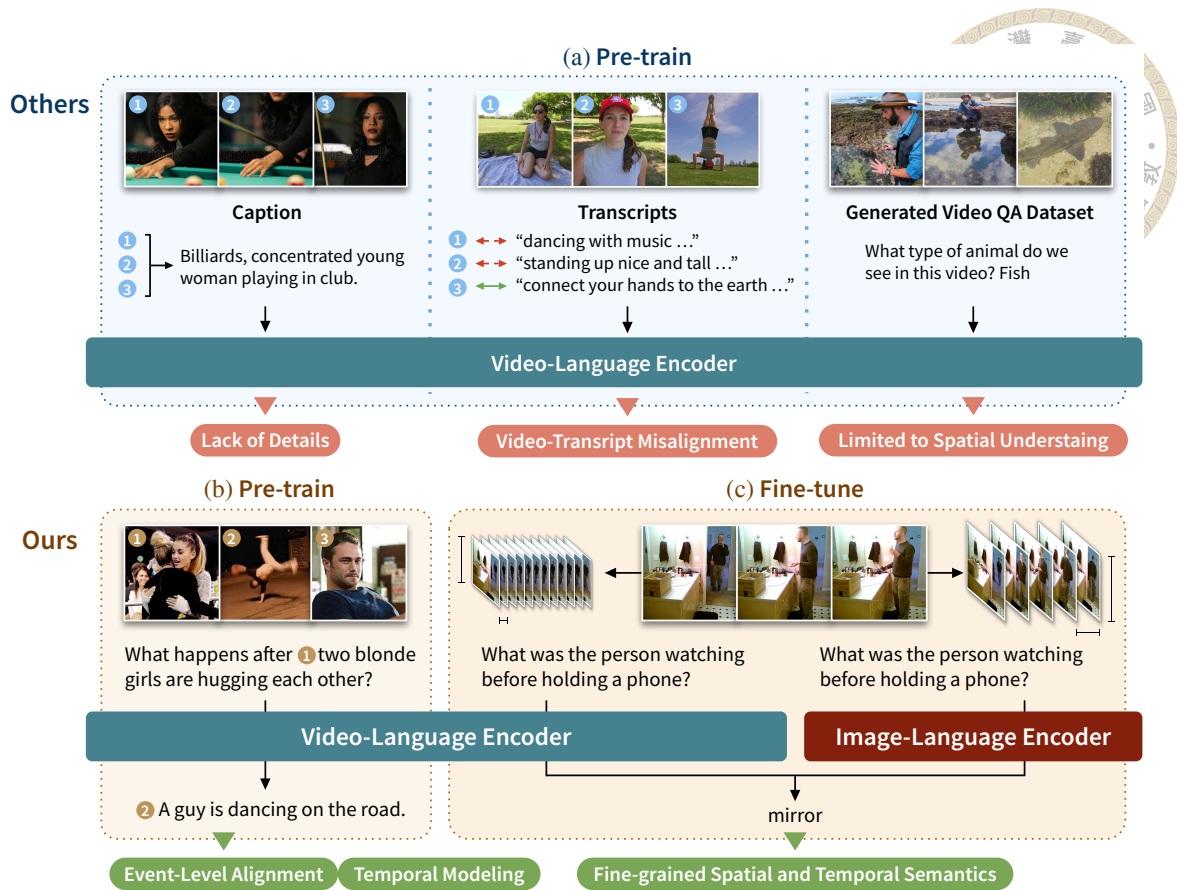
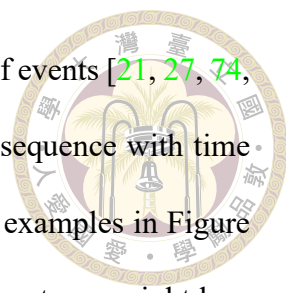


Figure 1.1: Comparison between (a) previous and (b)(c) our approaches for video QA. (a) Prior work solved video QA by video-language pre-training but might suffer from lack of event details, video-transcript misalignment or limited diversity of pre-training questions. (b) We pre-train a video-language encoder to learn event representations and temporal relations between them by asking the model to identify specific events in synthesized video sequences. (c) We integrate the video-language model with a pre-trained image-language model to encode fine-grained spatial and temporal semantics at different spatial-temporal resolutions.

tributed to the nature of video QA: the answers to the questions pertaining to only spatial semantics, without specifying time, are usually consistent across all related frames. This property suggests the potential of encoding fine-grained spatial semantics with only IL models.

In addition to spatial modeling, prior work modeled only coarse-grained temporal relations. A question involving temporal relations in video QA often refers to specific



events happening in periods of time and inquires about the sequence of events [21, 27, 74, 85]. It is thus essential to model events in videos and associate the sequence with time conjunctions in questions, such as *before* and *after*. However, as the examples in Figure 1.1 (a), prior approaches [18, 59, 60, 71, 91] aligning a video with a sentence might lose details of sequential events (what happens after the woman hit the ball), while matching short clips with transcripts [41, 86] may suffer from noise as spoken words often contain something not related to scenes [49]. Others [79, 80] pre-training on generated video QA datasets were mostly limited to spatial understanding. In fact, another examination reveals that the performance with shuffled frame inputs of some of these approaches is similar to that with normal inputs on video QA benchmarks requiring temporal modeling (see more details in Section 4.1.2). The result suggests developing a more effective strategy for modeling temporal relations.

To obtain fine-grained encoding of spatial and temporal semantics for video QA, we propose a novel pipeline decoupling spatial-temporal modeling into IL and VL encoders, illustrated in Figure 1.1 (c). With IL models adept at more fine-grained spatial modeling, we incorporate a pre-trained IL model to encode static spatial information independent of time from sparsely sampled frames at high spatial resolution. For questions requiring temporal relations, we train a VL encoder to model temporal dynamics, operating at high temporal but low spatial resolution. These two streams complement each other by paying attention to disparate aspects of videos.

To effectively model temporal relations for video QA, the VL encoder has to recognize events in videos, build their temporal relations, and associate such relations with languages containing temporal information. To this end, we introduce a novel pre-training objective, Temporal Referring Modeling (TRM). Depicted in Figure 1.1 (b), TRM queries

absolute and relative positions of events in videos synthesized by concatenating clips sampled from video captioning datasets [43, 73]. The concatenation simulates transitions of scenes and events in videos. Answering such queries requires a model to aggregate contiguous frames into events and distinguish adjacent events from distant ones. These operations help a model to learn both short- and long-term temporal dynamics.

We validate our model on two video QA benchmarks, ActivityNet-QA [85] and AGQA 2.0 [22]. The former contains diverse question types requiring spatial or temporal semantics, and the latter weaves spatial and temporal information together in each question to evaluate compositional reasoning. Our model outperforms the previous state-of-the-art. Ablation studies also demonstrate the efficacy of the proposed pipeline and pre-training objective.

In summary, we make the following key contributions. (i) With IL and VL models demonstrating complementary advantages, we decouple spatial and temporal modeling into a hybrid pipeline composed of both models to encode fine-grained visual semantics. (ii) We present a novel pre-training objective, Temporal Referring Modeling, to learn temporal relations between events by requesting models to identify specific events in video sequences. (iii) We outperform previous VL state-of-the-art methods on two benchmarks with orders of magnitudes less data for pre-training.



Chapter 2 Related Work

2.1 Video Question Answering

To encode, accumulate and build relationships between visual contents and between modalities for video QA, conventional approaches adopted Recurrent Neural Networks with attention [27, 77, 87, 89, 90], Memory Networks [13, 19, 31, 51, 68], Graph Neural Networks [23, 30, 44, 52, 54, 75], Modular Networks [34], and self-attention [29, 42, 69]. By pre-training large-scale VL datasets, Transformers [70] have further improved the interaction between modalities and made great progress in video QA [18, 41, 60, 71, 79, 80, 86, 91]. Our approach is built on the benefit of modeling relationships with pre-trained Transformers. In contrast to prior work, we carefully examine and take the individual advantage of IL and VL pre-training to encode spatial and temporal semantics.

2.2 Pre-training for Temporal Relation Modeling

VL pre-training learns to model temporal relationships via different approaches.



2.2.1 Learning from Global Alignment.

[18, 48, 60, 66, 71, 91] pre-trained models on datasets where a sentence delineates a single event of the entire corresponding video. With features of two modalities being aligned globally, events happening sequentially in a video are compressed, and details of events not mentioned in descriptions are likely lost. Such representations are not fine-grained enough for questions referring to specific moments.

2.2.2 Learning from Local Alignment and Frame Ordering.

[41, 86] pre-trained models over datasets with dense annotations such as video transcripts [50]. They matched segmented visual features with utterances and required models to order shuffled or any two frames. With this approach, models learn event-level but weak alignment between videos and languages as spoken words do not always correspond to visual contents [49]. Besides, ordering frames without grounding in languages make models learn, instead of temporal relations, rational predictions of what is likely to happen before and after an event, which is more related to visual common sense [1, 24, 53].

2.2.3 Learning from Large-Scale Video Question Answering Datasets.

[79, 80] pre-trained VL models over large-scale video QA datasets. The diversity of pre-training questions thus determines the effectiveness and capacity of transferred knowledge, but generated questions in [79] and [80] mainly pertain to scene and dialogue understanding, leaving temporal relationship modeling unsolved.

2.3 Encoding Motion and Appearance



Prior arts have explored two-stream networks to encode motion and appearance for action recognition [11, 15–17, 63, 72]. [10, 14] combined different spatial and temporal resolution to separately encode slow- and fast-changing scenes, and [57, 58] searched for multi-stream connectivity. Analogously, our two streams complement each other by focusing on disparate aspects of videos, but while their two streams both encode short-term actions, our IL stream aggregates scene information independent of time, and the VL stream encodes entire videos and constructs the temporal relationships between all actions and events.

Some recent work revealed that understanding of temporality is not always necessary to solve VL tasks. [35, 36] taking sparsely sampled frames outperformed previous methods. [4] provided stronger baselines with single frame inputs. However, with new tasks requiring temporal modeling proposed, such conclusions are likely to be circumscribed. We thus take a further step by proposing an effective strategy to encode fine-grained temporal semantics.





Chapter 3 Method

We introduce our video QA pipeline (Section 3.1) and the pre-training objective, Temporal Referring Modeling (Section 3.2). Implementation details are described in the supplement.

3.1 Decoupled Spatial-Temporal Encoders

The coarse-grained spatial modeling of prior approaches motivates us to develop more effective architectures, and IL models have shown great potential. While most VL models take scene or multi-frame features pre-extracted by image or action recognition models [41, 48, 79, 86], region features [47, 64, 67, 88] and features processed by attention [2, 78] have been proved powerful for IL models. These features provide detailed information of visual elements along with their spatial relations. As static scene information, if asked by questions without specifying time, are usually consistent across related frames, IL models should also be competent to encode fine-grained spatial relations for video QA.

Therefore, we propose a video QA pipeline that decouples spatial and temporal modeling by integrating an IL and a VL encoder. The IL encoder takes sparsely sampled frames at high spatial resolution as input. These frames are unordered and build consen-

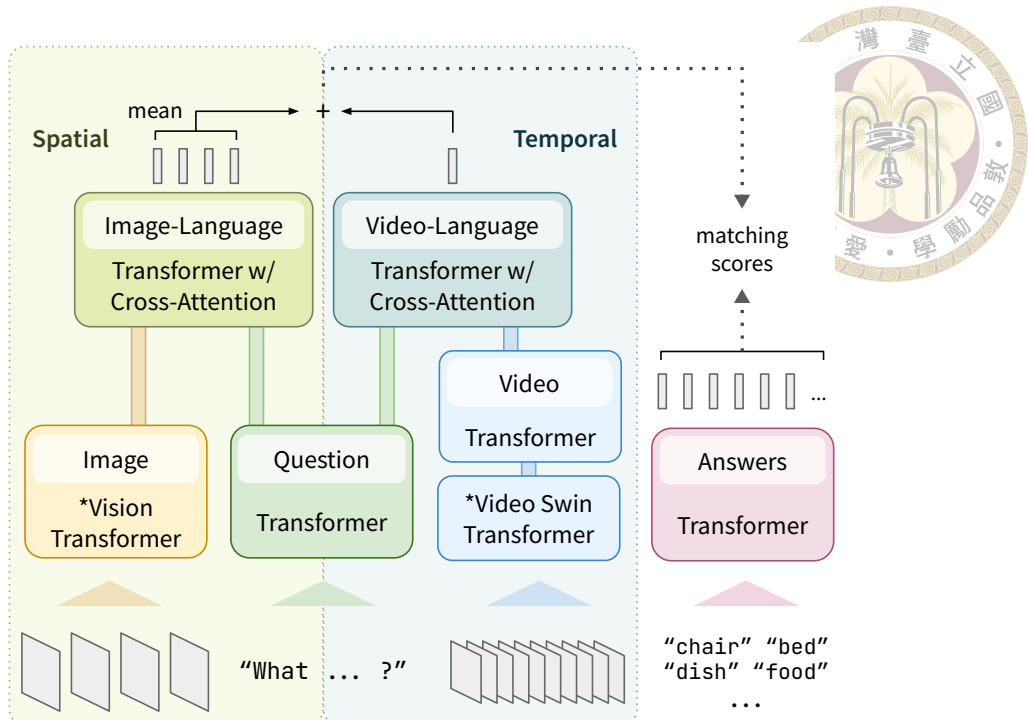


Figure 3.1: Our video QA pipeline. Encoded questions are fused with frames and videos to gather spatial and temporal information. Their representations are then compared with answer candidates to obtain the final predictions. (* marks the frozen modules.)

sus on fine-grained spatial information of static scenes. The VL encoder with input action features at low spatial but high temporal resolution recognizes and models the transitions of actions and events. These two streams of information are fused at the final stage to jointly form the prediction. We leave other ways of fusion for future exploration.

As illustrated in Figure 3.1, the pipeline consists of an image encoder, a video encoder, and a question encoder to process inputs, as well as an IL encoder and a VL encoder, both with cross-attention [26, 35, 39, 40], to perform multi-modality interaction. Another answer encoder encodes answer candidates, similar to [79]. To answer a question about a video, the question, video, and frames that are sparsely sampled from the video are encoded by their respective encoders. The question features then perform cross-attention to both frame and video features. The sum of two multi-modality representations is finally compared with encoded answer candidates to obtain the prediction. Formally, Q denotes

the input question. $\{\mathcal{I}^1, \dots, \mathcal{I}^T\}$ are T frames sampled from the input video \mathcal{V} , where $T \ll$ the length of \mathcal{V} . The question \mathcal{Q} is first encoded into a sequence of embeddings $\mathbf{w} = \{w_{\text{cls}}, w_1, \dots, w_L\}$, $w \in \mathbb{R}^D$, where w_{cls} is the embedding of the [CLS] token, and L is the number of word tokens. Then \mathbf{w} is fused with the frames and video as described below.

3.1.1 Image-Language Encoding.

For each t from 1 to T , the image encoder transforms frame \mathcal{I}^t into a sequence of patch embeddings $\mathbf{u} = \{u_{\text{cls}}^t, u_1^t, \dots, u_N^t\}$, $u \in \mathbb{R}^D$, where N is the number of patches. Then the question feature \mathbf{w} and frame feature \mathbf{u} are fused by the IL encoder with cross-attention and transform into $\{x_{\text{cls}}^t, x_1^t, \dots, x_L^t\}$, $x \in \mathbb{R}^D$. The multi-modality representation of the IL stream r is the average of [CLS] token embeddings x_{cls}^t of all frames encoded by a final multi-layer perceptron (MLP):

$$r = \frac{1}{T} \sum_{t=1}^T \text{MLP}(x_{\text{cls}}^t), r \in \mathbb{R}^D. \quad (3.1)$$

3.1.2 Video-Language Encoding.

The video feature extractor first encodes the input video \mathcal{V} into a sequence of features $\mathbf{e} = \{e_1, \dots, e_M\}$, $e \in \mathbb{R}^H$, where M is the length of the feature sequence. To indicate the beginning and the end of the video, we add two learnable tokens before and after the feature sequence. Temporal position encoding is also added to each feature to indicate the temporal order. Next, the feature sequence \mathbf{e} are contextualized and transformed into $\mathbf{v} = \{v_{\text{bos}}, v_1, \dots, v_M, v_{\text{eos}}\}$, $v \in \mathbb{R}^D$, where v_{bos} and v_{eos} are the beginning and the end

token after contextualization. The question feature \mathbf{w} then performs cross attention to the video feature \mathbf{v} through the VL encoder and transforms into $\{y_{\text{cls}}, y_1, \dots, y_L\}$, $y \in \mathbb{R}^D$. The multi-modality representation of the VL stream $s \in \mathbb{R}^D$ is the output of the first token y_{cls} transformed by a final MLP.



3.1.3 Answer Selection.

Following [79], another text encoder encodes the answer candidates (collected from all answers in training data with frequency > 1 for open-ended QA). The prediction of each candidate is the dot product between each encoded candidate and the sum of two multi-modality representations. Formally, \mathcal{A} denotes the answer set. For all $a \in \mathcal{A}$, we take the [CLS] token $z_{\text{cls}}^a \in \mathbb{R}^D$ of a 's feature. Then the logit of a is obtained via:

$$p_a = (r + s)^\top z_{\text{cls}}^a, p \in \mathbb{R}. \quad (3.2)$$

3.2 Temporal Referring Modeling

To pre-train the multi-modality encoders with affordable computation resources, we adopt an IL encoder pre-trained with image question answering (image QA), specifically VQA [20], and train the VL encoder for fine-grained temporal modeling with a novel objective.

Modeling fine-grained temporal relations for video QA requires the encoder to understand videos as event sequences and to associate the temporal relations of events with descriptions containing time conjunctions. To this end, we develop Temporal Referring Modeling (TRM), which, in the form of video QA, inquires about absolute and relative

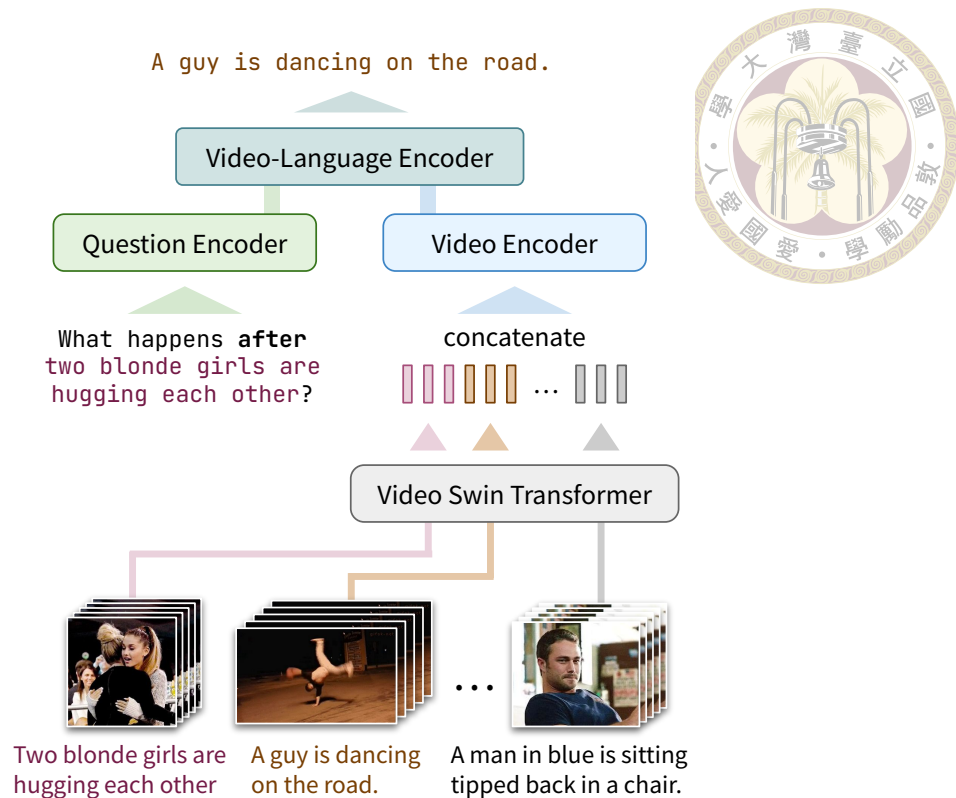


Figure 3.2: Temporal Referring Modeling, which associates visual events and their temporal relationships with languages by asking absolute and relative positions of events in concatenated video features sampled from video captioning data.

temporal positions of events in videos. As depicted in Figure 3.2, given a video composed of multiple events, TRM asks the model four questions: what happens at the beginning, at the end, before an event, or after an event. The model then selects an event description as the answer. To accomplish this task requires the model to identify events and manage the sequence.

TRM needs VL data which offers (1) event-level annotations that delineate scenes and events for segments of videos and (2) descriptions that explain the temporal dynamics of these segments. Ideally, dense video captioning [32] would be appropriate materials, but many of its time segments overlap, making the temporal relationships ambiguous. Labeling cost also hinders scalability. Thus, to satisfy the two conditions, we develop a simple yet effective way to generate data. As the example in Figure 3.2, we concate-

nate videos sampled from video captioning datasets to create videos with scene and event transitions. Then we generate questions by filling the question templates with captions of these videos. Incorrect answers are the other captions in the same video sequences, making the task more difficult.



Take, as an example, generating a video and a question that asks which event happens after an event. We first sample K pairs from a video captioning dataset, with each pair k composed of a video \mathcal{V}_k and a caption \mathcal{C}_k . The videos are encoded by the feature extractor into feature sequences $\{e_1^k, \dots, e_{M_k}^k\}$ for all k from 1 to K , where M_k is the length of features of \mathcal{V}_k . These sequences are then concatenated and form $\mathbf{e} = \{e_1^1, \dots, e_{M_1}^1, e_1^2, \dots, e_{M_K}^K\}$. To generate the question, we first sample a captions \mathcal{C}_i where $1 \leq i < K$, $i \in \mathbb{N}$. Then the question \mathcal{Q} is “What happens after \mathcal{C}_i ?” with the choices $\mathcal{A} = \{\mathcal{C}_k \mid 1 \leq k \leq K, k \neq i, k \in \mathbb{N}\}$ and the correct answer \mathcal{C}_{i+1} . Other questions are constructed similarly, where the answers to the questions about the beginning and the end are \mathcal{C}_0 and \mathcal{C}_K respectively. With all input the same as general video QA, the encoded feature \mathbf{w} of question \mathcal{Q} and the video feature \mathbf{e} are input to the VL encoder, going through the encoding and contextualizing process described in Section 3.1. The final objective is to minimize a standard cross-entropy loss.



Chapter 4 Experiments

We elaborate on the preliminary analysis of spatial and temporal reasoning capability of prior work (Section 4.1). Then we demonstrate the improvement in two video QA benchmarks with the proposed pipeline and Temporal Referring Modeling (Section 4.2). Ablation study is lastly presented evaluating the efficacy of each component. (Section 4.3).

4.1 Preliminary Analysis

Baselines. We take ALBEF [40] as an example of IL models. For VL models, we study VIOLET [18], HERO [41], and Just-Ask [79], which respectively instantiate three approaches discussed in Section 2.2. These are state-of-the-art of each approach with public code bases.

4.1.1 Encoding Spatial Semantics

We first assess the ability of encoding spatial semantics of IL models and VL models¹. ALBEF is run as image QA by sampling frames from a video and averaging frame predictions.

¹Just-Ask and VIOLET as HERO does not support open-ended QA

Type	Just-Ask	VIOLET	ALBEF	UB
Motion	28.00	18.25	32.50	70.63
Spatial Rel.	17.50	15.00	24.38	75.63
Temporal Rel.	4.88	2.12	3.75	32.88
Yes / No	66.28	71.87	79.75	100.00
Color	34.29	31.28	57.39	98.99
Object	26.73	22.33	31.45	70.13
Location	35.75	30.57	36.01	86.79
Number	50.17	50.33	55.61	99.83
Other	36.82	33.02	40.16	71.98
Overall	38.86	37.44	46.66	80.74



Table 4.1: Comparison between prior methods and our upper bound of ActivityNet-QA by question type. ALBEF exhibits advantages on the questions involving spatial reasoning. (Rel. is short for Relationships, and UB is the abbreviation for upper bound.)

Benchmark. We conduct the analysis on ActivietNet-QA [85], which contains 5.8K videos of human activities in daily life and 58K question-answer pairs spanning diverse categories across spatial and temporal semantics offering comprehensive evaluations.

Results. Table 4.1 contrasts the accuracy (acc) by question type of the IL model with other VL models. ALBEF, though without temporal modeling, is highly adept at spatial reasoning, such as Spatial Relationships and Color, while Just-Ask demonstrates a slight advantage in Temporal Relationships. Due to the removal of rare answers following [79], we report our performance upper bound of each type, which is the proportion of questions in the test set whose answers appeared in the training set. The tiny number of Temporal Relationships reveals the long-tail distribution of its answers, which partially explains the poor performance.

4.1.2 Modeling Temporal Relationships

We evaluate the capability of modeling temporal relationships by shuffling input frames and measuring the performance drop. Models are first trained with normal in-



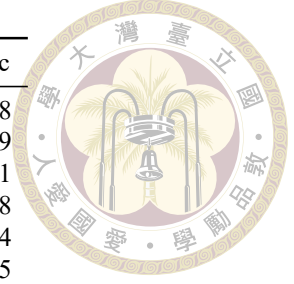
Method	Benchmark	Accuracy
VIOLET	AGQA	49.15
	AGQA*	49.22±.02
Just-Ask	AGQA	51.27
	AGQA*	47.73±.06
HERO	VIOLIN	69.01
	VIOLIN*	68.71±.08

Table 4.2: Results of prior work taking shuffled frames as input. The little performance drop indicates that some methods are not sensitive to the order of frames. (* signifies that input frames are shuffled. We report the average of three results for the shuffle experiment.)

put and tested their performance with shuffled input. Intuitively, taking shuffled frames as input should be detrimental to the performance of the questions requiring temporal modeling, such as those inquiring about the sequence of actions or events in videos.

Benchmarks. For VIOLET and Just-Ask, we conduct the study on AGQA 2.0 [22], a large-scale open-ended video QA benchmark where spatial and temporal information is required in each question for evaluating compositional reasoning. It contains 2.27M question-answer pairs and 9.6K videos. For HERO, we consider VIOLIN [45], a task of judging hypotheses from visual premises, which has been officially tested in their experiments.

Result. In Table 4.2, Just-Ask demonstrates the slight capability of temporal modeling, while VIOLET and HERO are not sensitive to the order of input frames, and their performances of taking normal and shuffled input frames are similar. The result suggests clear insufficiency for temporal relationship modeling.



Method	Pre-training Data	Acc
CoMVT [60]	100M	38.8
Just-Ask [79]	69M vid	38.9
MV-GPT [59]	100M	39.1
SiaSamRea [83]	5.6M img	39.8
MERLOT [86]	180M vid	41.4
VIOLET [18]	180M vid + 2.5M vid + 3M img	37.5
FrozenBiLM [81]	10M vid	43.2
Singularity [35]	14M img + 2.5M vid	44.1
Ours	14M img + 120K VQA + 14K vid	46.8

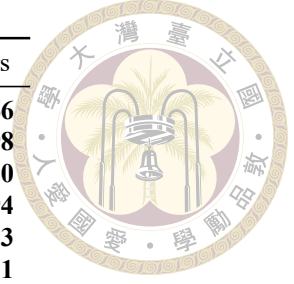
Table 4.3: Comparison with previous methods on ActivityNet-QA. We outperform all methods with significantly less pre-training data. The dataset names are provided in the supplement. (img: images. vid: videos.)

4.2 Video Question Answering

Table 4.3 compares our method with prior work on ActivityNet-QA. We outperform all previous methods with orders of magnitudes less pre-training data. The performance of each question type is listed in Table 4.4, where “Best” shows the highest scores among the three methods in Table 4.1, and “Diff” lists the difference between Best and our performance in proportion to Best. Our hybrid model performs comparably with the IL model in spatial modeling and boosts Temporal Relationships, verifying the efficacy of TRM.

Type	Best	Ours	Diff (%)
Motion	32.50	35.75	10.00
Spatial Rel.	24.38	23.88	-2.05
Temporal Rel.	4.88	5.25	7.58
Yes / No	79.75	78.61	-1.43
Color	57.39	59.11	3.00
Object	31.45	30.50	-3.02
Location	36.01	36.27	0.72
Number	55.61	55.28	-0.59
Other	40.16	39.63	-1.32
Overall	46.66	46.79	0.28

Table 4.4: Comparison with prior methods on ActivityNet-QA by question type. We perform comparably in question types of spatial information and improve temporal modeling.



	Type	Best w/o PT	Best w/ PT	Ours
Reasoning	Object-Rel.	40.33	48.91	59.66
	Rel.-Action	49.95	66.55	72.98
	Object-Action	50.00	68.78	75.20
	Superlative	33.55	39.83	48.94
	Sequencing	49.78	67.01	73.53
	Exists	50.01	59.35	63.21
	Duration Compar.	47.03	50.49	60.39
	Activity Recog.	5.52	21.53	27.78
Semantic	Object	40.40	49.31	61.27
	Rel.	49.99	59.60	63.93
	Action	47.58	58.03	65.96
Structure	Query	36.34	47.98	61.22
	Compare	49.71	65.11	72.04
	Choose	46.56	46.90	53.01
	Logic	50.02	56.20	59.18
	Verify	50.01	58.13	63.02
Overall	Binary	48.91	55.35	62.61
	Open	36.34	47.98	61.22
	All	42.11	51.27	61.91

Table 4.5: Comparison with prior work on AGQA 2.0. We list the best performance among methods without (Best w/o PT) and with pre-training (Best w/ PT) for each question type. Ours exceeds all methods in all question types. (Rel.: Relationships. Compar.: Comparison. Recog.: Recognition.)

Table 4.5 presents the performance on AGQA 2.0, which offers extensive annotation of multiple abilities necessary to answer each question. We list the highest accuracy among the methods without pre-training reported by [22] (“Best w/o PT”) and the higher scores between Just-Ask and VIOLET (“Best w/ PT”). Our method surpasses all prior work in all question types. The full table and detailed analysis are provided in the supplement.

4.3 Ablation Studies

We present the influence of input modalities and pre-training over AGQA 2.0 to study the effect of modeling decisions. As listed in Table 4.6, question-only input reveals the language bias, which serves as a baseline. The boost in performance with frames and



Question	Frames	Video	Acc
✓			41.32
✓	✓		50.07
✓	VQA		51.00
✓		✓	51.08
✓		TRM	55.62
✓	VQA	✓	56.61
✓	VQA	TRM*	56.97
✓	VQA	TRM	61.91

Table 4.6: Ablation study of input modalities and pre-training strategies on AGQA 2.0. The results favor our hybrid pipeline and TRM. (✓ means the modality is presented. VQA: pretrained on VQA. TRM: pre-trained with TRM. *: shuffled input.)

videos suggests successful encoding. Pretraining the IL encoder with VQA and the VL encoder with TRM both enhance the modeling capacity further. The performance drop due to shuffling videos verifies the efficacy of TRM. The full results are included in the supplement.

Stream	Acc
Image-Language	49.91
Video-Language	16.56
Both	61.91

Table 4.7: Ablation study of two encoding streams on AGQA 2.0.

In Table 4.7, we ablate the IL or VL stream. A model is trained with both streams and tested on AGQA 2.0 with a single stream. The performance drastically drops in both settings, proving that our hybrid model is not a trivial ensemble.



Chapter 5 Conclusion

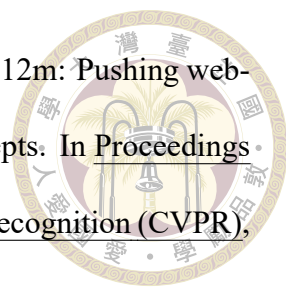
In this work, we propose decoupling spatial-temporal modeling by integrating IL and VL models to encode fine-grained visual semantics. Besides, by developing an objective that pre-trains the VL model to capture event-level temporal relations, we advance the visual understanding for video QA with much less pre-training data.





References

- [1] H. Agrawal, A. Chandrasekaran, D. Batra, D. Parikh, and M. Bansal. Sort story: Sorting jumbled images and captions into stories. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 925–931, Austin, Texas, Nov. 2016. Association for Computational Linguistics.
- [2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
- [3] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In IEEE International Conference on Computer Vision, 2021.
- [4] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles. Revisiting the "video" in video-language understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2917–2927, June 2022.
- [5] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600, 2018.

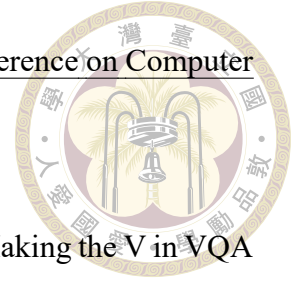
- 
- [6] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3558–3568, June 2021.
- [7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.
- [8] S. Choi, K.-W. On, Y.-J. Heo, A. Seo, Y. Jang, M. Lee, and B.-T. Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. Proceedings of the AAAI Conference on Artificial Intelligence, 35(2):1166–1174, May 2021.
- [9] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingular description of videos through latent topics and sparse object stitching. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 2634–2641, 2013.
- [10] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhof, and L. Van Gool. Large scale holistic video understanding. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, Computer Vision – ECCV 2020, pages 593–610, Cham, 2020. Springer International Publishing.
- [11] A. Diba, V. Sharma, and L. Van Gool. Deep temporal linear encoding networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby.

An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.



- [13] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
- [14] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
- [15] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc., 2016.
- [16] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
- [17] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
- [18] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu. Violet : End-to-end video-language transformers with masked visual-token modeling, 2021.
- [19] J. Gao, R. Ge, K. Chen, and R. Nevatia. Motion-appearance co-memory networks

for video question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.



[20] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[21] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[22] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala. Agqa 2.0: An updated benchmark for compositional spatio-temporal reasoning. arXiv preprint arXiv:2204.06105, 2022.

[23] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan. Location-aware graph convolutional networks for video question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11021–11028, 2020.

[24] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In AAAI, 2021.

[25] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In International Conference on Learning Representations, 2022.

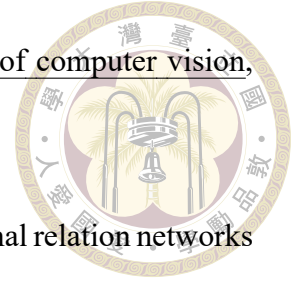
[26] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver:

General perception with iterative attention. In International conference on machine learning, pages 4651–4664. PMLR, 2021.



- [27] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
- [28] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 4904–4916. PMLR, 18–24 Jul 2021.
- [29] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):11101–11108, Apr. 2020.
- [30] P. Jiang and Y. Han. Reasoning with heterogeneous graph alignment for video question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):11109–11116, Apr. 2020.
- [31] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo. Progressive attention memory network for movie story question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
- [32] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In International Conference on Computer Vision (ICCV), 2017.
- [33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using

crowdsourced dense image annotations. International journal of computer vision, 123(1):32–73, 2017.



[34] T. M. Le, V. Le, S. Venkatesh, and T. Tran. Hierarchical conditional relation networks for multimodal video question answering. Int. J. Comput. Vision, 129(11):3027 – 3050, nov 2021.

[35] J. Lei, T. L. Berg, and M. Bansal. Revealing single frame bias for video-and-language learning, 2022.

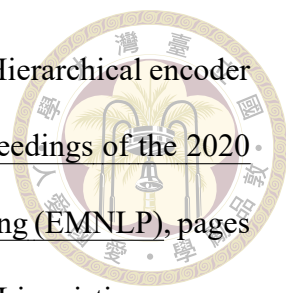
[36] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7331–7341, June 2021.


[37] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. In EMNLP, 2018.

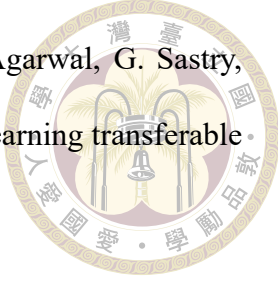
[38] J. Lei, L. Yu, T. Berg, and M. Bansal. What is more likely to happen next? video-and-language future event prediction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8769–8784, Online, Nov. 2020. Association for Computational Linguistics.

[39] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022.

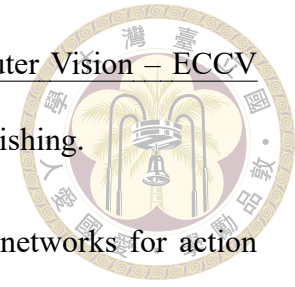
[40] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems, 2021.

- 
- [41] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2046–2065, Online, Nov. 2020. Association for Computational Linguistics.
- [42] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):8658–8665, Jul. 2019.
- [43] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. TGIF: A New Dataset and Benchmark on Animated GIF Description. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
- [44] F. Liu, J. Liu, W. Wang, and H. Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1698–1707, October 2021.
- [45] J. Liu, W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang, and J. Liu. Violin: A large-scale dataset for video-and-language inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
- [46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3202–3211, June 2022.
- [47] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

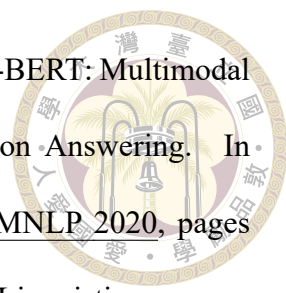
- 
- [48] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020.
- [49] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In CVPR, 2020.
- [50] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
- [51] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. In Proceedings of the IEEE International Conference on Computer Vision, pages 677–685, 2017.
- [52] J. Park, J. Lee, and K. Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15526–15535, June 2021.
- [53] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi. Visualcomet: Reasoning about the dynamic context of a still image. In In Proceedings of the European Conference on Computer Vision (ECCV), 2020.
- [54] L. Peng, S. Yang, Y. Bin, and G. Wang. Progressive Graph Attention Network for Video Question Answering, page 2871–2879. Association for Computing Machinery, New York, NY, USA, 2021.

- 
- [55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [56] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. Int. J. Comput. Vision, 123(1):94–120, may 2017.
- [57] M. S. Ryoo, A. Piergiovanni, M. Tan, and A. Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. In International Conference on Learning Representations, 2020.
- [58] M. S. Ryoo, A. J. Piergiovanni, J. Kangaspunta, and A. Angelova. Assemblenet+: Assembling modality representations via attention connections. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, Computer Vision – ECCV 2020, pages 654–671, Cham, 2020. Springer International Publishing.
- [59] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid. End-to-end generative pretraining for multimodal video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17959–17968, June 2022.
- [60] P. H. Seo, A. Nagrani, and C. Schmid. Look before you speak: Visually contextualized utterances. In Computer Vision and Pattern Recognition (CVPR), 2021.
- [61] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of ACL, 2018.
- [62] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In

B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision – ECCV 2016, pages 510–526, Cham, 2016. Springer International Publishing.



- [63] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.
- [64] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In International Conference on Learning Representations, 2020.
- [65] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Learning video representations using contrastive bidirectional transformer, 2019.
- [66] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
- [67] H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [68] M. Tapaswi, Y. Zhu, R. Stiefelwagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4631–4640, 2016.

- 
- [69] A. Urooj, A. Mazaheri, N. Da vitoria lobo, and M. Shah. MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4648–4660, Online, Nov. 2020. Association for Computational Linguistics.
- [70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.
- [71] A. J. Wang, Y. Ge, R. Yan, Y. Ge, X. Lin, G. Cai, J. Wu, Y. Shan, X. Qie, and M. Z. Shou. All in one: Exploring unified video-language pre-training, 2022.
- [72] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision – ECCV 2016, pages 20–36, Cham, 2016. Springer International Publishing.
- [73] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In The IEEE International Conference on Computer Vision (ICCV), October 2019.
- [74] J. Xiao, X. Shang, A. Yao, and T.-S. 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 (CVPR), pages 9777–9786, June 2021.
- [75] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T.-S. Chua. Video as conditional graph

hierarchy for multi-granular question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

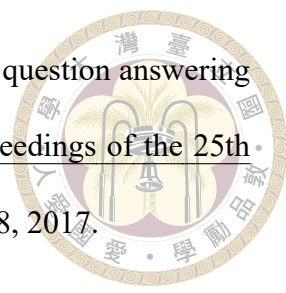


- [76] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM International Conference on Multimedia, MM '17, page 1645–1653, New York, NY, USA, 2017. Association for Computing Machinery.
- [77] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. 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.
- [78] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning - Volume 37, ICML'15, page 2048–2057. JMLR.org, 2015.
- [79] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Just ask: Learning to answer questions from millions of narrated videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1686–1697, 2021.
- [80] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Learning to answer visual questions from web videos. IEEE transactions on pattern analysis and machine intelligence, PP, 2022.
- [81] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Zero-shot video question answering via frozen bidirectional language models, 2022.
- [82] K. Yi*, C. Gan*, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. Clevrer:



Collision events for video representation and reasoning. In International Conference on Learning Representations, 2020.

- [83] W. Yu, H. Zheng, M. Li, L. Ji, L. Wu, N. Xiao, and N. Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 26462–26474. Curran Associates, Inc., 2021.
- [84] Y. Yu, J. Kim, and G. Kim. A joint sequence fusion model for video question answering and retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
- [85] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):9127–9134, Jul. 2019.
- [86] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. Merlot: Multimodal neural script knowledge models. In Advances in Neural Information Processing Systems 34, 2021.
- [87] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun. Leveraging video descriptions to learn video question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1), Feb. 2017.
- [88] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5579–5588, June 2021.

- 
- [89] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical dual-level attention network learning. In Proceedings of the 25th ACM international conference on Multimedia, pages 1050–1058, 2017.
- [90] Z. Zhao, Q. Yang, D. Cai, X. He, Y. Zhuang, Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In IJCAI, volume 2, page 8, 2017.
- [91] L. Zhu and Y. Yang. Actbert: Learning global-local video-text representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.



Appendix A — Implementation Details

A.1 Model Architectures

We introduce the details of modules in our video QA pipeline. Following [40] and [18], the image encoder is a 12-layer Vision Transformer [12], and the video encoder contains a Video Swin Transformer [46] (Swin-B) pre-trained on Kinetics-600 [5] for feature extraction and a 6-layer Transformer for contextualization. The question and answer encoder are both 6-layer Transformers [70] with each layer composed of a self-attention operation and a feed-forward network (FFN). The image- and video-language encoder are two 6-layer Transformers where each layer contains an additional cross-attention operation [25, 26, 35, 39, 40], in which text features serve as queries and perform attention to visual features. The question, image and image-language encoder are the same as the modules of ALBEF [40] pre-trained on VQA [20]. The video contextualization module and video-language encoder are initialized from the question and image-language encoder respectively. The image and video encoder are fixed during the whole training process. Detailed parameters are listed in Table A.1.

As the optimization of video encoding is not included in video-language training, we extract and save video features to save memory. We operate the Video Swin Transformer with the same configuration of Swin-B, which samples every two frames and transforms



Hyperparameter	Value
Embedding Size (D)	768
Number of Patches (N)	576
Video Feature Size (H)	1024
FFN Inner Hidden Size	3072
Number of Attention Heads	12
Attention Dropout	0.1
Dropout	0.1

Table A.1: Hyperparameters for the architecture.

a window of 32 frames into one feature. For long videos, for example ActivityNet [85] with an average of 180 seconds, we shift the window by 32 frames. For others, such as the datasets used in pre-training or AGQA 2.0 [22], we shift the window by 16 frames, and thus every window overlaps with the half of its previous and next window. Features of extremely long videos are sampled such that all videos are within a limited length.

A.2 Video-Language Pre-training

A.2.1 Details of Question and Video Synthesis for Temporal Referring Modeling

Temporal Referring Modeling (TRM) generates questions to inquire about absolute and relative temporal positions of specific events in videos. A question is formed by choosing from five templates and filling video descriptions into the template. The choice of templates includes “What happens?”, “What happens at the beginning?”, “What happens at the end?”, “What happens before [event x]?”, and “What happens after [event x]?”, where the first question is irrelevant to temporal relations but incorporated to facilitate video-language matching. The other four questions are designed for resemblance to video QA requiring temporal modeling, such as Temporal Relationships in ActivityNet-

QA [85] or State Transition in TGIF-QA [27].



Except for the first question paired with a single video, the corresponding videos of other questions are synthesized by concatenating videos sampled from video captioning datasets. This operation simulates a sequence of events that happen one after another and provides us the exact position of each event.

One may be concerned that the transitions of events in real videos are rather smooth and ambiguous, instead of clear difference between videos in a random concatenated video sequence where people, objects and almost the entire scenes drastically change. For example, in a video where people clean up the table after finishing dinner in the dining room, most of the visual elements, such as the people and scene, remain the same, but we humans can easily recognize these two events by comparing the actions and interaction of the people in the video. While TRM cannot generate such videos, our model has learned similar capability with TRM to compare human actions and interactions between moments. During fine-tuning, it can focus on adapting to smooth transitions, and thus learn faster than models with neither the capability of temporal reasoning nor event recognition.

A.2.2 Auxiliary Objective with Contrastive Learning

In addition to TRM, we apply an auxiliary objective during pre-training, which aligns video features with corresponding captions by contrastive learning, widely used in image- and video-language pre-training [28, 40, 48, 65, 71, 86]. Specifically, with the concatenated video feature sequence $\mathbf{e} = \{e_1^1, \dots, e_{M_1}^1, e_1^2, \dots, e_{M_K}^K\}$, we add the beginning and the end token before and after the sequence, as well as the temporal position encoding to each feature. Then after contextualization, we have $\mathbf{v} = \{v_{\text{bos}}, v_1^1, \dots, v_{M_1}^1, v_1^2, \dots, v_{M_K}^K, v_{\text{eos}}\}$.

To align each video to its caption, the objective learns a similarity function $\text{sim}(v, c) = g_v(f_v(v))^T g_c(f_c(c))$, such that parallel video-caption pairs have higher similarity scores. f_v produces the representation of \mathcal{V}_k , which averages the features of a video, e.g. $f_v(\mathcal{V}_k) = \sum_{m=1}^{M_k} v_m^k$, and f_c delivers the representation of a caption, which is the [CLS] embeddings of the caption feature encoded by the question encoder. g_v and g_c are two linear transformations that map the two representations into a normalized lower-dimensional space.

Following [40], we calculate the softmax-normalized video-to-caption and caption-to-video similarity as:

$$p_k^{\text{v2c}}(\mathcal{V}_k) = \frac{\exp(\text{sim}(\mathcal{V}_k, \mathcal{C}_k)/\tau)}{\sum_{i=1}^K \exp(\text{sim}(\mathcal{V}_k, \mathcal{C}_i)/\tau)}, \quad p_k^{\text{c2v}}(\mathcal{C}_k) = \frac{\exp(\text{sim}(\mathcal{C}_k, \mathcal{V}_k)/\tau)}{\sum_{i=1}^K \exp(\text{sim}(\mathcal{C}_k, \mathcal{V}_i)/\tau)}, \quad (\text{A.1})$$

where τ is a learnable temperature parameter. To increase the difficulty, we collect video-caption pairs from all video sequences in the same mini-batch B , and thus K is K times the size of a mini-batch in practice. Then, similar to [40, 55], let $\mathbf{y}^{\text{v2c}}(v)$ and $\mathbf{y}^{\text{c2v}}(c)$ denote the ground-truth one-hot similarity, where the probability of positive and negative pair are 1 and 0. The video-caption contrastive loss is defined as the cross-entropy CE between \mathbf{p} and \mathbf{y} :

$$\mathcal{L} = \frac{1}{2} \mathbb{E}_{(\mathcal{V}, \mathcal{C}) \sim B} [\text{CE}(\mathbf{y}^{\text{v2c}}(\mathcal{V}), \mathbf{p}^{\text{v2c}}(\mathcal{V})) + \text{CE}(\mathbf{y}^{\text{c2v}}(\mathcal{C}), \mathbf{p}^{\text{c2v}}(\mathcal{C}))] \quad (\text{A.2})$$

A.2.3 Pre-training Datasets

TRM samples video-caption pairs from video captioning datasets. We want the datasets as diverse as possible, not limited to cooking [9], movies [56], or indoor actions [62]. To maintain the computation within an affordable size, videos cannot be too long [32], or a video sequence would consist of few videos, which prohibit the model from

learning long-term temporal dependency.

We pre-train the video-language encoder over VATEX [73] and TGIF [43]. VATEX contains 41k videos from Kinetics-600 [5] and 826k sentences, where each video is paired with multiple descriptions. The length of the videos are all 10 seconds, cropped for precise action recognition in Kinetics. TGIF is an open-domain dataset containing 100K animated GIFs from Tumblr and 120K sentence descriptions. The duration of each GIF is around 3.1 seconds. We leave pre-training with longer videos and larger datasets for future work.



A.3 Optimization

Hyperparameter	Pre-Train	ActQA	AGQA
Learning Rate (Base)	1e-5	2e-5	2e-5
Learning Rate (Video)	5e-5	2e-4	5e-5
Learning Rate (MLP)	2.5e-4	1e-3	2e-4
Learning Rate (Ans)	2e-5	2e-5	2e-5
Weight Decay	1e-2	1e-2	1e-2
AdamW ϵ	1e-8	1e-8	1e-8
AdamW β_1	0.9	0.9	0.9
AdamW β_2	0.98	0.98	0.98
Training Steps	60K	-	-
Training Epochs	-	5	4
Warmup	0.03	0.1	0.1
Batch Size	128	64	64
Max Video Length	100	100	100
Max Question Length	50	-	-
Number of Videos (K)	8	-	-
Number of Frames (T)	-	16	8

Table A.2: Hyperparameters for pre-training (Pre-Train), ActivityNet-QA (ActQA), and AGQA 2.0 (AGQA). Base: the question, image, and image-language encoder. Video: the video and video-language encoder. Ans: the answer encoder.

The pre-training and fine-tuning are all optimized with AdamW optimizer and linear decay scheduling after warmup. All experiments are run with two NVIDIA RTX 3090s, with which the pre-training takes about 18 hours. Detailed hyperparameters are provided

in Table A.2.





Appendix B — Experiment Details

B.1 Details of Temporal Modeling Analysis

Some may question our preliminary analysis of temporal modeling, in which we first train a model with normal inputs and test it with normal and shuffled inputs. The performance drops imply the sensitivity to the order of frames, and thus little difference may indicate the incompetence of temporal modeling. Training and testing a model with shuffled input can also completely eliminates the temporal information, but this approach only reveals how well a model solves a task with spatial information (or dataset bias if the task is designed for evaluating temporal modeling), and thus it is not suitable for assessing a model’s capability of temporal modeling.

We conduct the analysis on AGQA and VIOLIN as some other video QA benchmarks are less appropriate. For example, some questions in ActivietNet-QA need only spatial knowledge. In NeXT-QA [74], while 29% questions are about temporal relations, others aim at spatial information or more advanced cognition, *e.g.* causal reasoning. The split of State Transition in TGIF-QA [27], though expected to fit this analysis well, could be solved by VIOLET without understanding the order of frames in our experiment (Table B.3).

Method	Benchmark	Accuracy
VIOLET	TGIF-QA	95.34
	TGIF-QA*	95.36 \pm .08



Table B.3: Results of VIOLET taking shuffled frames as input on the questions of State Transition of TGIF-QA. (* signifies that input frames are shuffled. We report the average of three results for the shuffle experiment.)

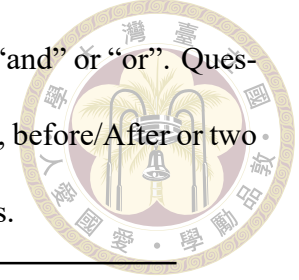
B.2 Pre-training Data Used by Prior Approaches

Compared with state-of-the-art approaches, our video QA pipeline achieves better performance on ActivityNet-QA with orders of magnitude less pre-training data. We list some widely-used pre-training datasets that are abbreviated in Table 3 of the main paper: 100M: HowTo100M [50]; 69M: HowToVQA69M [79]; 180M: YT-Temporal-180M [86]; 2.5M: WebVid [3]; 14M/3M: Conceptual Caption [6, 61]; 5.6M: COCO [7]+VisualGenome [33].

B.3 Full Results and Analysis on AGQA 2.0

AGQA 2.0 provides extensive annotations that each question is associated with reasoning abilities necessary to answer the question. The annotations cover four aspects: reasoning types, semantics class, structures, and answer types. Reasoning types define the design of question templates for evaluating certain reasoning abilities. We list some examples of question templates created by [21] in Table B.4 for the following analysis of our model’s behavior. The semantics class of a question describes its main subject: an object, relationship or action. Question structures include open questions (query), comparing attributes of two options (compare), choosing between two options (choose), yes/

no questions (verify), and understanding of logical operator, such as “and” or “or”. Questions have binary answer type restrict answer choices, such as Yes/No, before/After or two specified options, while many answers are possible to open questions.



Reasoning Type	Example of Template
Object-Relationship	What/Who/When/Where/How did they <rel> <object>?
Relationship-Action	Did they <relation> something before or after <action>?
Object-Action	Did they interact with <object> before or after <action>?
Superlative	What were they <action> first/last?
Sequencing	What did the person do after <action>?
Exists	Did/Does/Do <concept> occur?
Duration Comparison	Did they <action1> or <action2> for longer?
Activity Recognition	What does the person do before/after/while <action>?

Table B.4: Reasoning types and examples of their templates of AGQA 2.0.

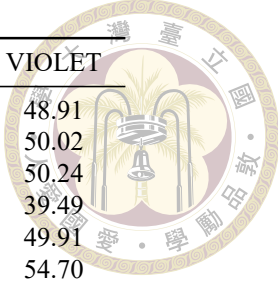
B.3.1 Full Results of Temporal Modeling Analysis

The full results of Table 2 in the main paper are presented in Table B.5, where we gauge the efficacy of temporal modeling of prior approaches by inputting shuffled videos and measuring performance drop. While Just-Ask [79] demonstrates improvement in Relationship-Action, Object-Action and Sequencing, VIOLET [18] performs similar in most types. The poor performance of VIOLET may be attributed to sparsely sampling, by which they enabled end-to-end training, but few frames seem not able to summarize the temporal dynamics of whole videos.

B.3.2 Full Results and Analysis of Our Method

We show the full results of our method on AGQA 2.0 with ablation of components and pre-training strategies in Table B.6.

We first examine the performance of inputting only questions (T), which reveals the



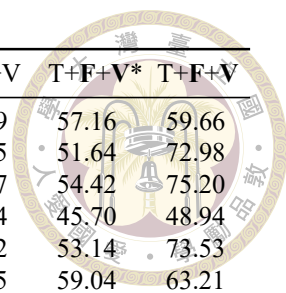
	Type	Just-Ask*	Just-Ask	VIOLET*	VIOLET
Reasoning	Object-Relationship	46.30	47.83	49.01	48.91
	Relationship-Action	50.78	66.55	50.04	50.02
	Object-Action	50.77	68.78	50.13	50.24
	Superlative	37.96	39.83	39.47	39.49
	Sequencing	50.66	67.01	49.86	49.91
	Exists	57.15	59.35	54.58	54.70
	Duration Comparison	50.66	50.49	30.70	30.64
	Activity Recognition	19.87	21.53	3.13	3.13
Semantic	Object	46.34	49.31	49.18	49.08
	Relationship	54.63	59.60	52.32	52.41
	Action	49.78	58.03	41.47	41.45
Structure	Query	45.53	47.25	48.15	47.98
	Compare	50.84	65.11	47.65	47.69
	Choose	39.78	41.00	46.97	46.90
	Logic	54.87	56.20	50.99	51.24
	Verify	56.22	58.13	55.42	55.46
Overall	Binary	49.95	55.35	50.30	50.33
	Open	45.53	47.25	48.15	47.98
	All	47.72	51.27	49.22	49.15

Table B.5: Full results of the preliminary analysis of temporal modeling on AGQA 2.0. (* means shuffled input. We report the result of one experiment.)

bias of the datasets as these questions can be solved without grounding to videos. With rigorous balancing procedure, this model cannot achieve more than 50% accuracy on any question type, but some questions, for example, those belonging to Relationship-Action, Object-Action, and Exists appear easier than others.

Inputting frames (T+F) improves the overall performance by about 10% accuracy, which mostly comes from Object-Relationship and Exists. It is reasonable as these questions involve less temporal information according to the templates, and they are more likely to be solved with a few static frames with spatial information about humans, objects, and scenes. Pre-training the image-language encoder with VQA [20] (T+F) shows further improvement in Exists, which seems more similar to the question design of image QA.

Accessing to videos (T+V) is helpful for different question types such as Superla-



Type	T	T+F	T+F	T+V	T+V	T+F+V	T+F+V*	T+F+V
Object-Relationship	39.15	49.21	50.33	51.67	53.40	56.39	57.16	59.66
Relationship-Action	50.05	50.61	50.00	49.83	71.57	53.25	51.64	72.98
Object-Action	49.99	50.11	50.00	50.03	74.74	56.27	54.42	75.20
Superlative	34.00	37.96	38.87	41.82	43.80	44.54	45.70	48.94
Sequencing	49.89	50.26	49.86	49.86	72.60	54.92	53.14	73.53
Exists	50.09	57.77	59.06	50.86	53.68	59.95	59.04	63.21
Duration Comparison	48.71	51.43	55.04	44.96	37.34	62.58	60.26	60.39
Activity Recognition	14.63	14.81	16.84	13.16	19.60	21.25	21.44	27.78
Object	39.25	49.24	50.16	51.44	55.28	56.50	57.31	61.27
Relationship	50.08	54.73	55.76	50.58	57.14	57.33	56.07	63.93
Action	48.49	49.98	50.86	47.09	56.52	56.35	54.39	65.96
Query	33.28	48.18	49.33	51.99	56.48	57.46	58.90	61.22
Compare	49.99	50.62	50.73	49.42	68.28	56.11	54.23	72.04
Choose	48.10	46.24	46.76	49.50	42.38	50.34	50.40	53.01
Logic	50.03	54.28	56.36	50.68	51.91	57.52	55.78	59.18
Verify	49.98	57.48	58.45	51.28	53.44	59.49	59.45	63.02
Binary	49.47	52.00	52.70	50.17	54.74	55.76	55.01	62.61
Open	33.28	48.18	49.33	51.99	56.48	57.46	58.90	61.22
All	41.32	50.07	51.00	51.08	55.62	56.61	56.97	61.91

Table B.6: Full results of our method on AGQA 2.0 with ablation of components and pre-training strategies. (T: questions; F: frames; **F**: frames with the image-language encoder pre-trained on VQA; V: videos; **V**: videos with the video-language encoder pre-trained with TRM; *: shuffled video inputs.)

tive, of which the questions ask about something happening first or last, but some other questions that also require temporal modeling, including Relationship-Action or Sequencing, are not improved. Besides, video inputs do not enhance the performance of questions improved by frame inputs. Such complementary advantages of frames and videos are consistent with our findings in the preliminary analysis, and inputting both frames and videos (T+F+V) does surpass inputting only one of them in all reasoning types.

Pre-training the video-language encoder with TRM (T+F+V) boosts the performance of most reasoning types, especially Relationship-Action, Object-Action, and Sequencing. These questions all need temporal modeling of event sequences in videos and have question formats more similar to TRM. The huge performance gap (20% accuracy) between normal (T+F+V) and shuffled video inputs (T+F+V*), as well as the little gap between no

pre-training (T+F+V) and shuffled inputs (T+F+V*), suggests successful temporal modeling and verifies the efficacy of TRM.



Despite the enhancement in most questions, TRM still struggles with some reasoning types, for example, Duration Comparison, which asks a machine which action lasts longer. These questions require a machine to memorize multiple events and identify their starting and ending point to obtain their duration. Such abilities are beyond the intention of developing TRM, and we leave it for future exploration.

B.3.3 Full Results of Ablation Study of Encoding Streams

Type	IL	VL
Object-Relationship	49.04	20.91
Relationship-Action	50.00	0.60
Object-Action	50.00	0.99
Superlative	36.00	15.86
Sequencing	49.85	0.80
Exists	57.09	0.07
Duration Comparison	58.99	0.00
Activity Recognition	13.06	0.92
Object	48.92	20.60
Relationship	54.58	0.20
Action	52.19	0.38
Query	47.35	26.68
Compare	51.24	0.69
Choose	48.29	22.11
Logic	55.09	0.04
Verify	56.51	0.08
Binary	52.52	6.30
Open	47.35	26.68
All	49.91	16.56

Table B.7: Full results of the ablation study on two encoding streams. (IL: image-language encoder; VL: video-language encoder.)

The full results of Table 7 in the main paper are reported in Table B.7, where we first train a model with both image- and video-language encoders, and test each stream with the test set. The image-language model demonstrates overwhelming advantages over

its video-language counterpart. However, this result cannot conclude the utility of any stream, for each stream can be trained to perform better than the question-only baseline. We hypothesize that temporal information can be seen as the complex evolution of spatial information, and thus when both streams cooperate in spatial-temporal modeling, the image-language stream offers overall understanding of visual elements and scenes, while the video-language stream assists it and models the detailed changes.

