Can I Trust Your Answer? Visually Grounded Video Question Answering

Junbin Xiao   Angela Yao*   Yicong Li   Tat-Seng Chua

Department of Computer Science, National University of Singapore

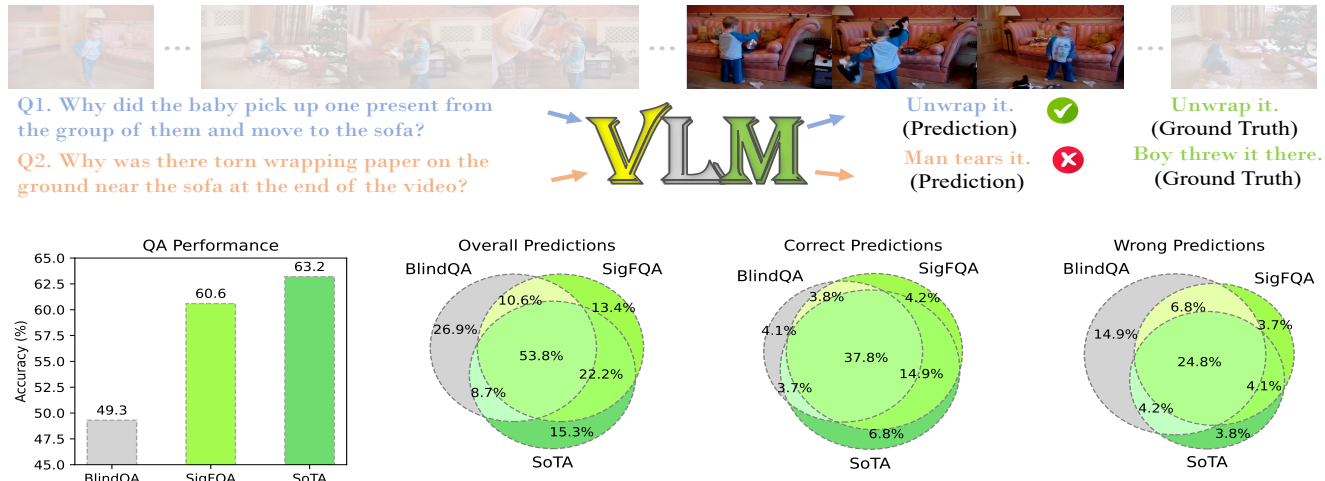{junbin,ayao,chuats}@comp.nus.edu.sg, liyicong@u.nus.edu

Figure 1. **Top**: Real predictions of VQA models (BlindQA, SigFQA and SoTA) on NExT-QA [55]. All the models correctly answer Q1 but wrongly answer Q2, albeit the two questions sharing visual evidence (*the boy unwraps the present and throws the wrapping paper*). **Bottom**: Overlap in model predictions. BlindQA: A pure language model (*i.e.* RoBERTa [36]) fine-tuned with question-answer text. SigFQA: An image-text model (*i.e.* CLIP [44]) using only the center video frame. SoTA: Temp[CLIP] (Sec. 5) model using 32 video frames. The analyses indicate that the models may not learn from causal visual content but more likely from language short-cut and irrelevant visual context.

Abstract

We study visually grounded VideoQA in response to the emerging trends of utilizing pretraining techniques for video-language understanding. Specifically, by forcing vision-language models (VLMs) to answer questions and simultaneously provide visual evidence, we seek to ascertain the extent to which the predictions of such techniques are genuinely anchored in relevant video content, versus spurious correlations from language or irrelevant visual context. Towards this, we construct NExT-GQA – an extension of NExT-QA with 10.5K temporal grounding (or location) labels tied to the original QA pairs. With NExT-GQA, we scrutinize a series of state-of-the-art VLMs. Through post-hoc attention analysis, we find that these models are extremely weak in substantiating the answers despite their strong QA performance. This exposes the limitation of current VLMs in making reliable predictions. As a remedy, we further explore and propose a grounded-QA method via Gaussian mask optimization and cross-modal learning. Experiments with different backbones demonstrate that this grounding mechanism improves both grounding and QA. With these efforts, we aim to push towards trustworthy VLMs in VQA systems. Our dataset and code are available at `https://github.com/doc-doc/NExT-GQA`.

1. Introduction

Video Question Answering (VideoQA) has recently emerged as a golden testbed to develop vision-language models (VLMs), especially foundation VLMs pretrained at scale on multi-modal web corpora [1,11,24,29,51,61,63,66]. Despite significant advancements in QA performance, a fundamental concern arises – whether or to what extent are the answers of such techniques grounded on the relevant visual content? Alternatively, are they relying on the *language short-cut* for the use of powerful language models [20,34,42,52,62,64,69,70,73] or *spurious vision-language correlation* captured via cross-modal pretraining [45,60]?

For example, Fig. 1(Top) shows that existing VLMs are inclined to answer questions with *language-biased* pre-

dictions, *e.g.*, "`unwrap (Q1: present)`" and "`tear (Q2: paper)`". Fig. 1(Bottom) shows that the overall predictions of SoTA VLMs overlap the predictions of standalone language models (BlindQA), *i.e.*, models without visual inputs, by 62.5%. In fact, the BlindQA counterpart shares 66% of correct predictions and also 79% of the wrong predictions in SoTA VLMs. The overlap increases by injecting a coarse visual signal from a single frame [3, 23] into the language model; as our later analysis will show, this frame often lies outside the key moments of the correct answers.

Given these findings, a natural question arises – to what extent are the predictions of current VLMs grounded on the video content, and more precisely on the *relevant* parts? To answer this, we propose to study **visually grounded VideoQA**. Grounded VQA requires VLMs to answer the questions and simultaneously output the relevant video moments to support the answers. Earlier works have explored grounded QA under full supervision [25, 26], but we target visual explanability in VideoQA and thus define the task under weak-supervision, which is the first of its kind.

To accomplish the goal, we construct the NExT-GQA (short for Grounded) dataset by extending the NExT-QA dataset [55] with $10.5K$ temporal labels of start and end timestamps for the QA pairs in the validation and test sets. The labels are manually annotated and checked to be key for comprehending the questions and determining the correct answers. With NExT-GQA, we examine a series of recent high-performing VLMs, including task-specific architectures without pretraining [57] and pretrained models with either image-text or video-text data [11, 44] and those using frozen large language models (LLMs) [62, 64]. Our findings reveal that all these models struggle to predict visually grounded answers, despite their strong QA performance. For example, the SoTA model [62] achieves QA accuracy of 69%, but only 16% of the correctly predicted answers are grounded in the video. In contrast, humans can ground 82% out of the 93% of the correctly answered questions. Such clear discrepancy underscores the need for continued research efforts.

As a pioneering solution, we propose a temporal grounding approach which can be easily applied to existing VLMs for visually grounded VideoQA. Specifically, our approach learns differentiable Gaussian masks along the temporal dimension of the videos, by optimizing light-weight transformer layers under both VQA and question-video (QV) supervisions, without the need for temporal labels. Experiments with different QA backbones demonstrate that our approach effectively improves video grounding and question answering as well. The improvement is especially significant on a subset of questions that necessitate video understanding and temporal grounding.

To summarize our contributions: 1) we conduct the first study of weakly grounded VideoQA, and release the NExT-GQA benchmark, to facilitate research on more trustworthy

VLMs; 2) we comprehensively analyze a wide range of advanced VLMs and reveal their limitation in performing visually grounded QA; 3) we propose a simple yet effective grounding mechanism which not only enhances existing VLMs in visual grounding but also contributes to new SoTA QA performance, *e.g.*, 73.1% on NExT-QA test set.

2. Related Work

Benchmarks Grounded VQA with full supervision has been studied in both image [4, 77] and video [25, 26] domains. Recently, weakly-supervised grounding has received increasing attention in ImageQA [18, 19] and video grounding [12, 39]. Nonetheless, to our best knowledge, there is no work for weakly-grounded VideoQA. Also, existing supervised benchmarks are either biased towards localizing subtitles in TV shows (*e.g.* TVQA [25]) or limited to few objects (*e.g.* VidSTG [72]). Thus, they are not ideal benchmarks for visual evidence grounding.

Techniques Strong **VideoQA** methods are predominantly banked on transformer [49] and pre-training [44]. The popular transformer architectures follow either shared [24, 50], dual [57, 58, 60, 61] or stacked [11, 29, 62] implementations, and pre-training is done with image-text [24], video-text [29, 61, 65, 66] or both [11] forms of data. Notably, all these VLMs use powerful language models (*e.g.*, BERT [8], T5 [45], GPT [2], LLaMA [48] or their successors) for text encoding and focus on improving QA while ignoring visual evidence grounding. Some recent works [5, 30–33, 43, 64] have begun to ground key frames or objects for VideoQA. Yet, they still aim to improve QA accuracy, and thus the grounded contents may not be the *actual* evidences since they do not evaluate grounding. For weakly-supervised **video grounding**, typical approaches extract temporal proposals and rank the proposals according to their similarities with the language query [13, 35, 39, 71]. Despite their effectiveness, these two-stage approaches are notorious for inefficient and sub-optimal for multi-granular temporal modelling. More recent research [74, 75] points to the superiority of end-to-end Gaussian mask learning. In light of this, we design a simple yet effective Gaussian mask learning module for grounding in VideoQA. Unlike previous works [74, 75] that design small transformer models and hand-craft negative visual proposals for contrastive learning, we integrate Gaussian mask learning into large VLMs and optimize its parameters via question-answering and video-question grounding.

Language Priors Our work is also related to efforts in preventing language priors and other spurious correlations. Goyal *et al*. [14] construct VQAv2 to prevent language priors in VQA by pairing the questions with additional images that carry similar contents but with different answers. Niu *et al*. [40] and Guo *et al*. [15] alleviate language priors by regularizing the prediction scores. Zheng *et al*. [67] develop

Table 1. Statistics of NExT-GQA.

Split	#Vid.	#Que.	#Seg.	Seg. Dur.(s)	Vid. Dur.(s)	Ratio (S./V.)
Train	3,860	34,132	-	-	44.9	-
Val	567	3,358	3,931	7.3	42.2	0.2
Test	990	5,553	6,600	6.7	39.5	0.2

X2-VLM via multi-grained vision-language pretraining for better spatial grounding. These works discourage short-cut learning in the image side by either collecting new data, designing tailored-made learning methods, or focusing on spatial grounding. Our primary contribution lies in defining the weakly-grounded VideoQA task to encourage more interpretable and trustworthy techniques.

3. NExT-GQA Dataset

3.1. Dataset Construction and Analysis

Data Source. We choose NExT-QA [55] as our data source to augment with temporal labels. Most of the other VideoQA datasets [17, 59] are not suitable because they feature short videos (3 ∼ 15s) already trimmed around the relevant content. NExT-QA has three different types of questions: Causal ("why/how"), Temporal ("before/when/after") and Descriptive ("what/who/where"). We exclude the descriptive questions because they mostly pertain to global video content (*e.g.*, "what event?") or their answers can be found almost throughout the whole video (*e.g.*, "where is?"). In addition, we only label the validation and test sets since we aim for a weakly-supervised setup. As a result, 11,378 QA pairs drawn from 1,570 videos [47] are to be annotated.

Label Collection. We invite undergraduate students for annotation (using Elan [10]) and train them with our demo annotations together with some trial examples following specific criteria (see Appendix A.1) before the actual annotation exercise. To guarantee quality and reduce subjectiveness, each QA pair is annotated by at least two people. The final temporal label is determined by an additional check and refinement of the two accepted annotations. The entire exercise lasted around 2 months with a team of 30 annotators. Eventually, we collect 10,531 valid temporal segments corresponding to 8,911 QA pairs and 1,557 videos. Detailed statistics are presented in Tab. 1.

Label Analysis. Fig. 2a(left) shows that most of the segments last for less than 15s, and with an average duration of 7s (Tab. 1) which is short compared to the video length (∼40s). In fact, the ratio reflected in Fig. 2a(right) shows that most of the segments occupy less than half (0.5) length of the videos, and the average ratio is merely 0.2 (Tab. 1. This ratio is slightly low, compared to that of 0.3 for both ActivityNet-Caption [21] and Charades-STA [12]. Moreover, Fig. 2b(1) shows that the segments are evenly distributed in the left, middle and right parts of the video. Fig. 2b(2) shows that near 90% of the QAs ground on a single temporal
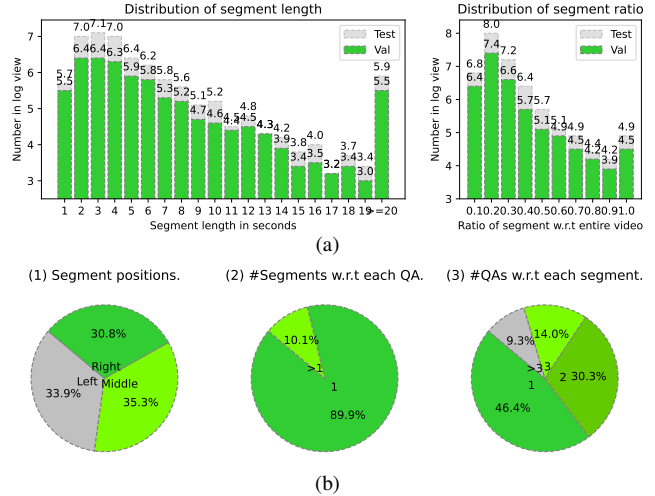

Figure 2. Distribution of temporal segments.

segment. Conversely, Fig. 2b(3) shows that each segment often corresponds to 1 or 2 QAs (Here two segments are considered the same if their IoU > 0.5). To better understand the dataset, we show two examples in Appendix Fig. 6.

3.2. Comparison with Existing Benchmarks

We highlight the uniqueness of NExT-GQA by comparing it with other relevant benchmarks in Tab. 2.

NExT-GQA *vs*. NExT-QA. NExT-QA [55] targets the prediction of text answers. NExT-GQA differs in two major aspects: 1) it provides visual evidence to support the answers, and 2) it extends the VQA setup by allowing visual answers. This satisfies more real-world applications and additionally helps to better diagnose model performance. For example, is a prediction wrong because the model failed to localize the relevant video contents, or because it could not convert the localized video contents into a text answer? NExT-GQA is also more challenging, because: 1) the models need to achieve multiple goals (*i.e.*, grounding and QA) and *maintain their consistency*, and 2) the set of questions are relatively harder to answer by focusing on local video moments in untrimmed long videos. This also differs from major VideoQA benchmarks that focus on trimmed (short) video understanding [17, 53, 59].

NExT-GQA *vs*. Video Grounding (VG) Benchmarks. Video grounding [12, 21] aims to find a video moment described by a declarative sentence. NExT-GQA shares core challenges, *i.e.*, *cross-modal correspondence learning* and *multi-granular temporal modelling*, while featuring some unique aspects. **First**, the questions feature visual content for grounding which is not explicitly stated in the text, such as "a baby falls and cries." *vs*. "why did the baby cry?". To answer the questions, the models not only need to find the described video moments (*e.g.*, "baby cries") but also should be capable of refining the moment

Table 2. Benchmark comparison. GD: Grounding. MM: Multi-modal. Acc: Accuracy. IoU/P: Intersection over Union/Prediction.

Datasets	GD	QA	Weak Sup.	Goal	Eval
ActNet-Cap [21]	√	×	√	VG	IoU
Cha-STA [12]	√	×	√	VG	IoU
TVQA [25]	√	√	×	MMVQA	Acc, IoU
VidSTG [72]	√	√	×	VG	IoU
NExT-QA [55]	×	√	×	VQA	Acc
NExT-GQA	√	√	√	Trust VQA	Acc, IoP, IoU

to enclose the answer (*e.g.*, "baby falls"). This may ask for *temporal and causal relationship* reasoning. **Second**, the video backgrounds are relatively monotonous with little scene change. Accordingly, the temporal segments of QA pairs are often more fine-grained than those in VG benchmarks. **Notably**, NExT-GQA prioritizes finding visual evidence to support the answers. This means that any individual frame or moment that sufficiently tells the answer should be considered as a valid grounding instead of retrieving all of the video contents that match the query. This is reflected in our selection of intersection over prediction (**IoP**) as an evaluation criterion. That is, a correct grounding depends on whether the predicted segment falls into the labelled segment but is not necessarily an exact match.

NExT-GQA *vs*. Supervised benchmarks. Fully-supervised benchmarks [25, 26] provide temporal annotations for training data; the labels can resolve reference ambiguities in the questions or improve QA performance with well-localized visual inputs. NExT-GQA differs from them by seeking to identify visual evidence that explains the answers with QA supervision alone. It is worth mentioning that directly applying the fully-supervised benchmarks for weakly grounding does not suit our goal, because these benchmarks are either biased to text localization [25] or the answers are a limited set of, *e.g.* 80 objects [72]. Additionally, we focus on weakly-supervised *temporal* grounding and leave *spatio-temporal* grounding for future exploration. Our consideration is that fine-grained *spatio-temporal* grounding [72] is currently more challenging than question-answering, especially in the weak supervision setting [54], and would derail the main goal of VQA.

4. Weakly-Supervised Grounding in VideoQA

VideoQA. We first give an overview of the typical approaches to VideoQA, focusing on transformer-based methods due to their superior performance. Given a video v and a question q, the goal of VideoQA is to predict a correct answer a^* from a set of candidate answers A. Depending on the task setting, A can be given by multiple choices accompanying each question [55, 56] (multi-choice), or by a global answer set [59] to all questions (open-ended). Note that SoTA transformer-methods [11, 57, 61, 62] formulate and solve both multi-choice QA and open-ended QA in a unified formulation:

$$a^* = \text{argmax}_{a \in A} \ \Psi(a|v, q, A), \qquad (1)$$

in which the mapping Ψ is typically realized as either shared [24, 50], stacked [11, 29, 62] or dual [44, 57, 61] transformer. In this work, we primarily study the behaviour of the stacked-(Fig. 3a) and dual-transformer (Fig. 3b) architectures for their relatively better performance.

Weakly Grounded VideoQA. Aside from answering questions, weakly-grounded VideoQA requires the models to explicitly estimate a QA-relevant video segment to serve as visual evidence. We introduce below three model-agnostic solutions to achieve this goal:

Post-hoc (PH). Intuitively, relevant temporal segments can be found through a *post-hoc* analysis of the temporal attention, *i.e.*, identifying the segment or frame with the maximal attention value and then thresholding around it to obtain a time interval. To that end, we use *attention-pooling* to summarize the outputs from the temporal transformers for dual architectures. For stacked architectures, we directly return the averaged multi-head attention values corresponding to the prediction token.

Naive Gaussian (NG). The *post-hoc* approach is designed to analyze the models, but not influence their predictions. More favourably, we propose to explicitly incorporate a video grounding mechanism into VideoQA. We illustrate the framework in Fig. 4a, and reformulate Eqn. 1 as

$$a^*, \ t^* = \text{argmax}_{a \in A} \ \Psi(a|v_t, q, A)\Phi(t|v, q), \qquad (2)$$

in which the grounding module Φ firstly estimates the key moment specified by t and thereafter the QA module Ψ takes the more localized video content v_t for answer prediction. To enable end-to-end learning, t is represented by differentiable Gaussian weights over the entire video sequence, *i.e.*, $t \sim N(\mu, \sigma^2)$, where $\mu, \sigma \in [0, 1]$ are two learnable Gaussian parameters corresponding to the mean and standard deviation. During inference, the grounding can be achieved by the confidence interval $t = (\mu - \gamma\sigma, \mu + \gamma\sigma) * d$, where γ is a hyper-parameter to control the width of the confidence interval and d denotes the duration of the video.

Fig. 3c shows a dual transformer instantiation of this naive solution. The difference with respect to the original VideoQA counterpart (Fig. 3b) lies in a Gaussian mask prediction head, along with a Gaussian weighted token learning and aggregation stage (details in Appendix A.2). We find that this approach effectively learns and outputs grounding information. Nevertheless, the improvements over a *post-hoc* solution are limited due to the weak QA supervision.

NG+. In light of the naive Gaussian results, we further design an auxiliary objective with cross-modal self-supervision to regularize the VQA objective towards more visually grounded QA. Specifically, for each question q^+,
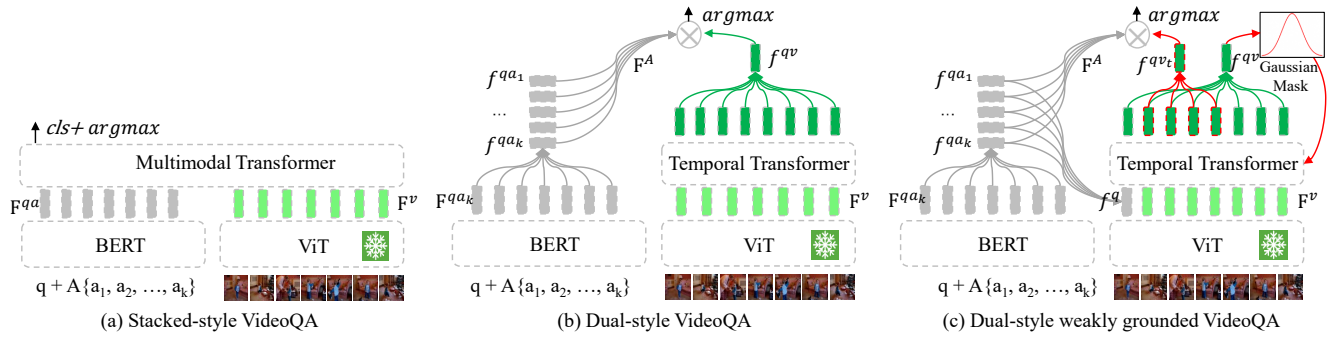
Figure 3. Illustration of stacked (a) and dual (b) style Transformer architecture for VideoQA. (c) Our example of dual-style weakly-grounded VideoQA. Note that the grounding part is identical for stacked-style implementation.

(a) Stacked-style VideoQA
(b) Dual-style VideoQA
(c) Dual-style weakly grounded VideoQA

(a) Framework of weakly-grounded VideoQA.
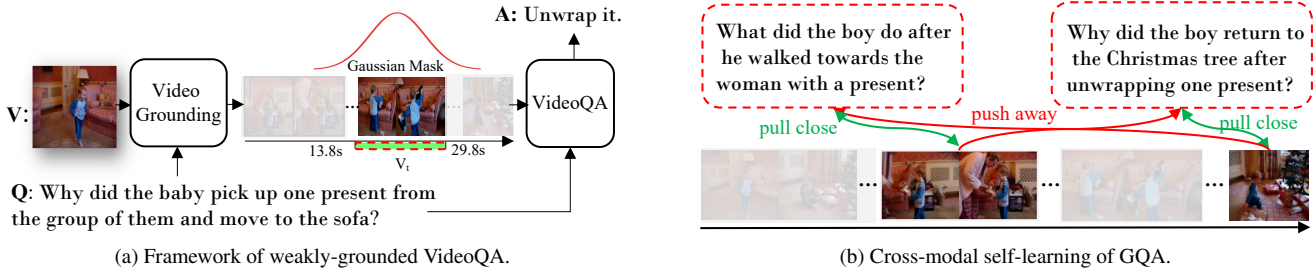
(b) Cross-modal self-learning of GQA.

Figure 4. Illustration of the framework (a) and our NG+ solution (b) for weakly-grounded VideoQA.

we treat the corresponding grounding hypothesis v_t as an anchor point and pull it close to q^+ while pushing it away from other questions Q^- in the feature space. The negative set Q^- includes: 1) other questions defined in the same video as *hard* negatives, since a large portion (near half) of questions each invokes unique video moment for answer (Fig. 2b(3)); 2) questions sampled from other videos to ensure the sufficiency and diversity of negative samples. Moreover, we enrich 10% of the positive questions by rephrasing each question (using GPT-4 [41]) with a maximum of 5 additional questions to form Q^+. It is worth noting that there is only one positive question at each training iteration and the enriched positive questions are randomly picked to substitute the original one for data augmentation. Such form of contrast is thus implemented as classification by also fixing the number of negative questions to be identical to that of distractor answers. Thereby, our final solution is:

$$a^*,\ t^* = \underbrace{\operatorname*{argmax}_{a \in A} \Psi(a|v_t, q^+, A)\Phi(t|v, q^+)}_{\text{GroundedQA}} +$$
$$\alpha \underbrace{\operatorname*{argmax}_{q \in Q} \Psi(q^+|v_t, Q)\Phi(t|v, q^+)}_{\text{Grounding}}, \qquad (3)$$

where $Q = Q^+ \cup Q^-$ which comprises both the positive and negative questions of v_t and α is a trade-off parameter. Note that the Grounding-term coarsely identifies the question-relevant video moment t, while the GroundedQA-term not only makes the prediction but also helps to refine the moment t with answer supervision. The overall objective

thus enforces the grounded video contents to be relevant to both the answers and the questions.

5. Experiments

5.1. Overview

Our experiments answer three research questions: **Q1**: To what extent are the current VLMs' predictions grounded on *relevant* video content? **Q2**: Does better QA performance imply better grounding and vice versa? **Q3**: How effective is our Gaussian masking mechanism? We study a wide variety of VLMs, covering different architectures (dual and stacked transformers), vision encoders (task-specific and pretrained with image- or video-text data), and text encoders (BERT, RoBERTa, DeBERTa, Flan-T5):

1. **VGT** [57] is a task-specific, dual-style graph transformer model. It encodes spatio-temporal object information [46] for VideoQA. We also investigate VGT with RoBERTa [36] as suggested by [58].

2. **Temp[Swin]** is a dual architecture. The Swin Transformer (SWT) [37] is pre-trained on ImageNet [7]. **Temp[CLIP]** and **Temp[BLIP]** follow the same dual architecture, but use ViT [9] pretrained by CLIP [44] and BLIP [28] respectively as vision encoders.

3. **VIOLETv2** [11] adopts a stacked transformer. It uses video Swin Transformer [38] (VSWT) and BERT for vision and text encoding, respectively. The model is

pretrained with both image- and video-text data, and achieves SoTA on various VL tasks.

4. **FrozenBiLM** [62] applies a stacked transformer. It uses CLIP as vision encoder and highlights the strength of adapting *frozen* large language models (LLMs) (*e.g.*, DeBERTa-V2-XL (1B) [16]) for VideoQA.

5. **IGV** [31] and **SeViLA** [64] are additionally reproduced for comparison. Both works emphasize grounding keyframes for VideoQA. IGV is built on a visual graph, whereas SeViLA is founded on BLIP-2 [27]. It exploits ViT-G [68] and *frozen* LLM (*e.g.*, Flan-T5-XL (3B) [6]) for video localization and QA. In our implementation, we choose the smallest time spans that can enclose the localized keyframes as grounded moments.

Experimental Settings. For all the models, we uniformly sample 32 frames from each video and freeze the vision encoders. In *post-hoc* analysis, the temporal attention thresholds are set dynamically according to mean attention values to maximize the grounded QA accuracy. The number of negative questions in Eqn. 3 is kept the same as the number of distractor answers in MCQA to facilitate joint optimization. The trade-off parameter α is set to 1 and 0.1 for dual and stacked transformers, respectively. During inference, the hyperparameter γ for the Gaussian confidence interval is chose from $\{1, 0.8\}$ depending on different models. Our final results are reported based on a combination of predictions from Gaussian and temporal attention. All hyper-parameters are tuned on the validation set, and unless otherwise indicated, the results are reported on the test set. Other details are described in Appendix A.2.

Evaluation. We report accuracy for QA [76], which stands for the percentage of correctly answered questions. For visual evidence grounding, we use intersection over prediction (IoP) to measure whether the predicted temporal window lies inside the ground truth. Additionally, we include temporal IoU following video grounding benchmarks. For both IoP and IoU, we report the mean and the values with overlap thresholds of 0.3 and 0.5. If a QA pair involves multiple temporal segments, we report the results based on the one with maximal overlap with the prediction. Notably, we define grounded QA accuracy (Acc@GQA) to inspect the percentages of questions that are correctly answered and also visually grounded (*i.e.*, IoP ≥ 0.5).

5.2. Result and Analysis

5.2.1 Q1:Are the answers visually grounded?

We focus on Acc@QA, Acc@GQA and IoP@0.5 in the *Post-hoc* (PH) block of Tab. 3. Generally, the existing VLMs excel at QA but are weak in grounding the answers in the videos. For example, all the methods exceed 50% in QA accuracy,

yet cannot reach more than 12-16% for grounded QA accuracy. In fact, the SoTA QA model (FrozenBiLM) achieves an accuracy of 69% for QA compared to a surprisingly low 16% for GQA. The results of IoP@0.5 suggest that the large disparity is mainly due to the models' poor performance in temporal grounding. It is also due partly to inconsistency between grounding and QA because not all correct grounding yields correct answers according to Acc@GQA *vs*. IoP@0.5. We additionally exclude the influence of sparse video sampling by investigating the coverage of QA content w.r.t the number of sampled video frames in Fig. 5(a). The figure shows that the sampled 32 frames can cover almost all QA contents. Moreover, to understand the extent of such poor performance, we estimate the upper-bound performance via a human study on 10% of the test data. The study shows that the participants correctly answered 93% of the questions, with 82% being visually grounded.

Given the above observations, we believe most of these models' answers are not grounded on the relevant video content. Instead, they are more likely derived from language shortcuts or spurious correlations with irrelevant visual context. **To investigate the language shortcut**, we conduct a BlindQA experiment, in which we train only the language counterparts of the VQA models without video inputs. Tab. 4(a) shows that BlindQA achieves 80% of the performance of standard VQA (NormalQA), *i.e.*, 50.3% *vs*. 59.4% for the dual models and 56.7% *vs*. 69.1% for the stacked models. **To study the spurious correlations**, we test the VLMs by directly sampling inside (PosQA) or outside (NegQA) the ground-truth video segments. Surprisingly, the models' QA performances remain almost unaffected compared to a normal uniform sampling (NormalQA), likely because the image representations are not fine-grained enough to differentiate different frames. Tab. 4(a) shows that providing the ground-truth temporal segments (PosQA) brings marginal improvement ($<1\%$) for the dual-style models and even hurts stacked-style transformers, likely due to a distribution shift in visual inputs. Furthermore, excluding the temporal segments (NegQA) degenerates the performance by less than 1% for both dual and stacked-style models. The above studies reinforce our belief that the current VLM's predictions are often not visually-grounded.

5.2.2 Q2: Does better QA imply better grounding?

First, by focusing on Acc@QA, mIoP and mIoU in Tab. 3. we find that **better QA is not necessarily established by better grounding, and the results vary across architectures**. For instance, by comparing across different architectures, FrozenBiLM shows the strongest QA performance, yet with surprisingly poor grounding, *e.g.*, the IoP values are even worse than those of VGT which displays the lowest QA results among other transformer models. This could be due to FrozenBiLM's freezing of the LLMs, causing its predic-

Table 3. Grounded QA performance on NExT-GQA test set. †: original NExT-QA. D/S: Dual/Stacked. CM: Cross-modal pretrain. BT: BERT. RBT: RoBERTa. DBT: DeBERTa-V2-XL. FT5: Flan-T5-XL. Random: always choose the same answer id and return the whole video duration as grounding result. *: pretrain on video-language grounding dataset.

	Model	D/S	CM	Vision	Text	Acc@QA	Acc@QA†	Acc@GQA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
	Human	-	-	-	-	93.3	-	82.1	72.1	91.7	86.2	61.2	86.9	70.3
	Random	-	-	-	-	20.0	20.0	1.7	21.1	20.6	8.7	21.1	20.6	8.7
	IGV	-	N	ResNet	BT	50.1	51.3	10.2	21.4	26.9	18.9	14.0	19.8	9.6
	*SeViLA**	S	Y	ViT-G	FT5	*68.1*	*71.5*	*16.6*	***29.5***	***34.7***	*22.9*	***21.7***	***29.2***	***13.8***
PH	VGT	D	N	RCNN	BT	50.9	53.8	12.7	24.7	26.0	24.6	3.0	4.2	1.4
	VIOLETv2	S	Y	VSWT	BT	52.9	57.2	12.8	23.6	25.1	23.3	3.1	4.3	1.3
	VGT	D	N	RCNN	RBT	55.7	57.7	14.4	25.3	26.4	25.3	3.0	3.6	1.7
	Temp[Swin]	D	N	SWT	RBT	55.9	58.7	13.5	23.1	24.7	23.0	4.9	6.6	2.3
	Temp[CLIP]	D	Y	ViT-B	RBT	57.9	60.7	14.7	24.1	26.2	24.1	6.1	8.3	3.7
	Temp[BLIP]	D	Y	ViT-B	RBT	58.5	61.5	14.9	25.0	27.8	25.3	6.9	10.0	4.5
	Temp[CLIP]	D	Y	ViT-L	RBT	59.4	62.5	15.2	25.4	28.2	25.5	6.6	9.3	4.1
	FrozenBiLM	S	Y	ViT-L	DBT	69.1	71.8	15.8	22.7	25.8	22.1	7.1	10.0	4.4
NG	Temp[CLIP]	D	Y	ViT-L	RBT	59.4	62.7	15.5	25.8	28.8	**25.9**	7.7	10.9	4.6
	FrozenBiLM	S	Y	ViT-L	DBT	70.4	73.1	17.2	24.0	28.5	23.5	9.2	13.0	5.8
NG+	Temp[CLIP]	D	Y	ViT-L	RBT	60.2+0.8	63.3+0.8	16.0+0.8	25.7+0.3	31.4+3.2	25.5+0.0	12.1+5.5	17.5+8.2	8.9+4.8
	FrozenBiLM	S	Y	ViT-L	DBT	**70.8**+1.7	**73.1**+1.4	**17.5**+1.7	24.2+1.5	28.5+2.7	23.7+1.6	9.6+2.5	13.5+3.5	6.1+1.7

Table 4. Performances under different settings. (+): with NG+. VQA: Question subset that BlindQA cannot answer. GDQA: Subset that both BlindQA and NegQA cannot answer but PosQA can.

(a)

	Model	NormalQA	BlindQA	PosQA	NegQA
Post-hoc	Temp[CLIP]	59.4	50.3	59.8	59.1
	FrozenBiLM	69.1	56.7	68.5	68.2
NG+	Temp[CLIP]	60.2	50.3	61.0	59.4
	FrozenBiLM	70.8	56.7	70.0	69.6

(b)

Models	QA Set	Acc@QA	Acc@GQA	mIoP	mIoU
Temp[CLIP]	Whole	59.4	15.2	25.5	6.6
	VQA	35.7	9.7	25.2	7.0
	VQA(+)	39.4+3.7	10.6+0.9	25.5+0.3	12.2+5.2
	GDQA	23.0	10.8	27.6	5.9
	GDQA(+)	30.2+7.2	14.3+3.5	29.3+1.7	13.1+7.2
FrozenBiLM	Whole	69.1	15.8	22.7	7.1
	VQA	47.6	11.2	22.2	6.6
	VQA(+)	50.0+2.4	12.8+1.6	23.7+1.5	9.5+2.9
	GDQA	42.6	14.8	24.6	7.3
	GDQA(+)	44.0+1.4	16.6+1.8	27.0+2.4	13.2+5.9

tions to heavily rely on the *common sense knowledge* of the LLMs rather than the provided videos (similar problem is also found on SeViLA). In contrast, VGT is a task specific model. It focuses on exploiting the fine-grained video information, and thus conditions better on the visual content. By comparing among different instantiations of the same architectures (*e.g.*, Temp[Swin] to Temp[CLIP]) as well as different training epochs of the same models in Fig. 5(b), we find that **the grounding performance (mIoP) improves along with the increase of QA accuracy for dual-style architectures yet not for stacked-style ones**. Second, regarding the influence of grounding on QA, our conclusion is that **having grounding is better than not having it**. Yet, this is not controlled and opts for the underlying shortcuts

when the models are allowed to learn freely. The conclusion is backed by the observations that PosQA always outperforms NegQA in Tab. 4(a) regardless of model architectures. Moreover, our effort to improve grounding also brings better QA performance (Tab. 3 & 4(b)). However, as mentioned, **correct grounding does not guarantee correct answers**.

5.2.3 Q3: Is Gaussian masking solution effective?

We incorporate our Gaussian grounding mechanism (NG and NG+) into the top-performing dual- and stacked-style models and compare with *Post-hoc* baseline. [1] Tab. 3 shows that both *NG* and *NG+* lead to better grounding and QA performance. Also, NG+ generally outperforms NG, especially for dual-style architectures. Additionally, Tab. 4(b) indicates that our superiority gets enlarged in answering the subset of questions that necessitate videos and temporal grounding.

For better understanding, we analyze two cases in Fig. 5(c). The top example shows that the Gaussian masks (*NG* and *NG+*) are more focused on the relevant video moment than temporal attention, thus bringing better grounding, especially for IoU. The bottom example highlights the strength of NG+. In this case, there are multiple visual instances that correspond to the answer "girl stands up". The correct instance is the one after the "girl takes the green ball", though the instance after "take the red ball" is more salient. Both the *Post-hoc* and *Naive* methods are distracted because they are learned via answer supervision alone. In contrast, *NG+* finds the correct grounding since it also optimizes the cross-modal correspondence between questions and video segments. More detailed analyses are presented in Appendix A.3.

---

[1] Despite the weaker performance, we highlight the higher efficiency of dual-style implementation, especially in retrieval-based QA systems as exemplified by multi-choice QA.
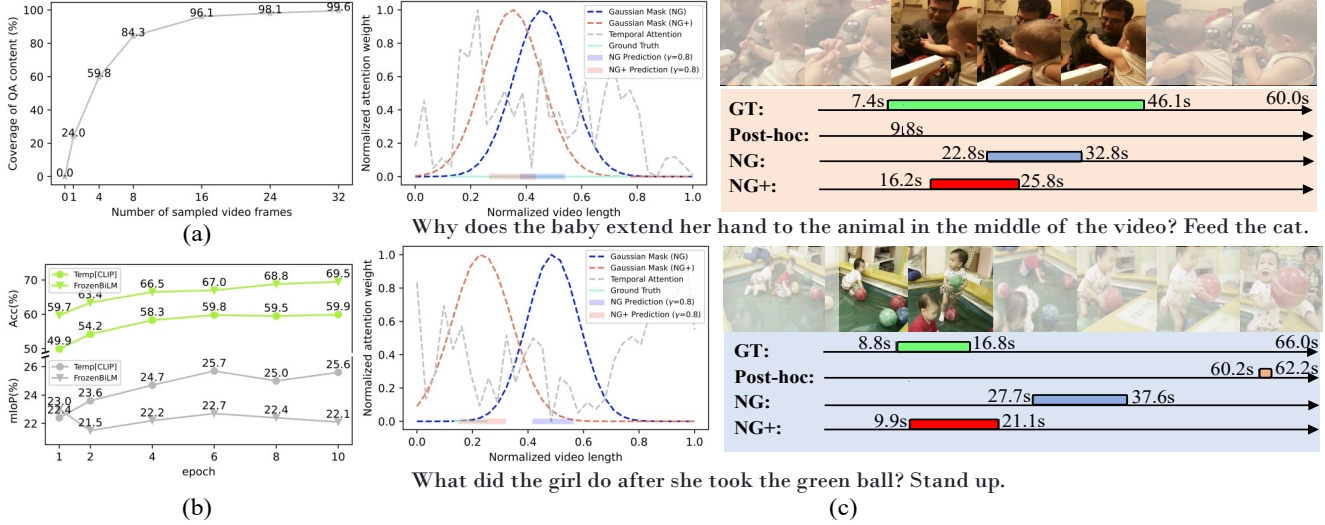
Figure 5. Analysis of visually-grounded VideoQA. (a) Coverage of QA content w.r.t. number of sampled video frames. (b) VQA and VG results w.r.t. training epochs on NExT-GQA Val set. (c) Visualization of the prediction examples. (Please zoom in for better view.)

5.2.4 Method Comparison

Compared with a random baseline, all methods effectively perform grounded QA (refer to Acc@GQA and IoP@0.5 in Tab. 3). More concretely, we find that both IGV and SeViLA obtain lower GQA accuracy than FrozenGQA though they also incorporate a sense of grounding in their models. The weakness manifest in both visual evidence grounding (IoP@0.5) and QA. However, we find that SeViLA performs much better than other methods in standalone grounding (mIoP and mIoU). We speculate this is because SeViLA is pretrained with localization supervisions [22]. The observations thus point to possible future improvement by pretraining with location supervisions. Furthermore, they call for improved coordination between QA and grounding.

5.2.5 Other Observations

Tab. 3 also compares the Acc@QA performance on NExT-GQA versus the full (original) NExT-QA test set. There is a consistent 2-3% higher accuracy on the full set, suggesting that the questions rooted in local video moments are harder to answer than those rely on overall video content. Besides, the **cross-modal pretrained representations perform better than the uni-modal pretrained ones** for both VQA and visual grounding. Also, the image-text pretrained representations outperform those pretrained with video-text data. Moreover, **existing dual-style architectures tend to have better grounding performance than stacked ones** (Note that FrozenBiLM's high-ranking GQA result is due to its strong QA performance but not grounding). This is surprising, as there is no cross-modal interaction in dual-style implementations. We speculate that cross-modal transformers likely suffer from a *uni-modal bias*, which leads to the attention being skewed towards the language side for

predicting textual answers. The findings on the one hand consolidate the benefits of harnessing foundation VLMs or LLMs for videoQA. On the other hand, they accentuate the need to balance between vision fact and text knowledge.

6. Conclusion

We summarize the following points and raise them as open challenges for the rest of the community: **First**, current VLMs built on powerful language models excel in answering visual questions. Yet, their predictions often lack a strong connection to the pertinent visual information but instead heavily rely on languages short-cut and irrelevant visual context. This calls for more efforts towards the interpretability and trustability. **Second**, our experiments show that, localizing the questions, especially those featuring temporal actions and events is still a difficult and open challenge. Our studies indicate that solving this problem would largely benefit visually-grounded VideoQA. **Third**, although our solution improves grounding and QA, there is still a large gap compared with human performance. This leaves ample opportunity for follow-up works. **Last** but not least, we highlight the significance of NExT-GQA and hope it can contribute towards advancement in these areas.

Limitations. The NG+ method demands more memory and time to train (Appendix A.3.3). Besides, our analyses are focused on multi-choice QA (Appendix A.4).

References

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

[2] 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. *NeurIPS*, 33:1877–1901, 2020.

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *CVPR*, pages 2917–2927, 2022.

[4] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *CVPR*, pages 19098–19107, 2022.

[5] Junwen Chen and Yu Kong Golisano. Explainable video entailment with grounded visual evidence. In *ICCV*, pages 2001–2010, 2021.

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[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. In *ICLR*, 2021.

[10] Nijmegen: Max Planck Institute for Psycholinguistics. Elan (version 6.5) [computer software]. (2023). the language archive. 2023.

[11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *CVPR*, 2023.

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[13] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. In *EMNLP*, pages 1481–1487, 2019.

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[15] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *TIP*, 31:227–238, 2021.

[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021.

[17] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017.

[18] Aisha Urooj Khan, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *CVPR*, pages 8461–8470. IEEE, 2021.

[19] Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Vitoria Lobo, and Mubarak Shah. Weakly supervised grounding for vqa in vision-language transformers. In *ECCV*, pages 652–670. Springer, 2022.

[20] Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*, pages 4300–4316, 2023.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[22] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34:11846–11858, 2021.

[23] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

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

[25] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[26] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, pages 8211–8225, 2020.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

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

[29] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020.

[30] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. In *ACM MM*, pages 4714–4722, 2022.

[31] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, 2022.

[32] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Transformer-empowered invariant grounding for video question answering. *T-PAMI*, pages 1–12, 2023.

[33] Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Discovering spatio-temporal rationales for video question answering. In *ICCV*, pages 13869–13878, 2023.

[34] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[35] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, volume 34, pages 11539–11546, 2020.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[38] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022.

[39] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601, 2019.

[40] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021.

[41] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[42] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *ECCV*, pages 76–94. Springer, 2022.

[43] Tianwen Qian, Ran Cui, Jingjing Chen, Pai Peng, Xiaowei Guo, and Yu-Gang Jiang. Locate before answering: Answer guided question localization for video question answering. *IEEE Transactions on Multimedia*, 2023.

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

[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. *JMLR*, 21(1):5485–5551, 2020.

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[47] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, pages 279–287, 2019.

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[50] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022.

[51] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[52] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *NeurIPS*, 2023.

[53] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021.

[54] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *ECCV*, pages 447–464. Springer, 2020.

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

[56] 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 *AAAI*, volume 36, pages 2804–2812, 2022.

[57] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, pages 39–58. Springer, 2022.

[58] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat Seng Chua. Contrastive video question answering via video graph transformer. *IEEE T-PAMI*, 45(11):13265–13280, nov 2023.

[59] 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 *ACM MM*, pages 1645–1653, 2017.

[60] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800, 2021.

[61] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021.

[62] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 35:124–141, 2022.

[63] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *ICCV*, 2023.

[64] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, 37, 2023.

[65] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, pages 16375–16387, 2022.

[66] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 34:23634–23651, 2021.

[67] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X2-vlm: All-in-one pre-trained model for vision-language tasks. *T-PAMI*, 2023.

[68] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, pages 12104–12113, 2022.

[69] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *EMNLP*, 2023.

[70] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[71] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, 33:18123–18134, 2020.

[72] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, pages 10668–10677, 2020.

[73] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, pages 6586–6597, 2023.

[74] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, volume 36, pages 3517–3525, 2022.

[75] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, pages 15555–15564, 2022.

[76] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *EMNLP*, pages 6439–6455, 2022.

[77] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016.