A. Appendix

A.1. Dataset Construction

Criteria. We set clear criteria to limit the ambiguities and subjectiveness. 1) For each question, the annotation should encompass the entire temporal segment which features the answer and also sufficient context to interpret the question. 2) If the visual content mentioned in the question is not simultaneous or contiguous in time with the answer, then the annotation should focus on the answer. 3) If visual evidence for an answer appears multiple times in the video, then all relevant video moments (individual segment) should be annotated. 4) If the answer of the question can be seen throughout the entire video, the question is omitted. Yet, to ensure we can collect sufficient labels, we pay annotators on a per annotated segment basis. Fig. 6 shows two examples of our annotation outcome.

A.2. Implementation Details

Post-hoc. For dual-style transformer, we have tried both *attention-pooling* the visual tokens as well as *prepending* a summary token and then averaging the multi-head attention of transformer. We find that the two methods bring similar QA performance. Yet, the *prepending* approach demands much more training epochs and thus we chose attention-pooling as our final solution. Moreover, to obtain a reasonable time span from the learned temporal attention, we treat the frame of maximal attention value as the pivot location, and search around it to enclose the frames whose attention values satisfying certain criteria. Before that, the attention values are normalized to [0, 1] using min-max method. The criteria of whether a frame should be enclosed are jointly determined by its attention score and its distance with the pivot frame. In our implementation, we also smooth the attention values and the distance threshold is set to 10s. Finally, the minimal frame id and the maximal frame id are mapped to the time seconds to obtain the temporal span. Note that the frame of maximal attention will always be selected.

Naive Gaussian (NG). For both dual- and stacked-style architectures, the Gaussian prediction head is implemented with a lightweight transformer layer followed by linear projectors. Specifically, the Gaussian mask G (with dimension equal to the length of frames sequence F) is propagated to each self-attention head to weight the original self-attention weights before aggregating (summarizing) the value vectors, *i.e.*, $F_h = G \cdot \text{softmax}(\frac{F^K(F^Q)^\top}{\sqrt{d_k}})F^V$, in which Q, K, V indicate the respective query, key and value vector in self-attention. Notably, as there is no independent visual stream in stacked-style transformer, we pick the tokens belonging to the visual inputs and go through the Gaussian-weighted transformer. The resultant tokens are then preprended back into the multi-modal token sequence for answer prediction.

Video-Question Correspondence Learning (NG+). We find that a two-stage training paradigm to pretrain with the Grounding-term and then finetune with both objectives in Eqn. 3 brings better performance than one-stage training. In both stage, the negative questions are selected from the same videos as the positive question at a chance of 0.3. Note that we exclude the descriptive questions because their answers usually appear throughout the video. Also at a chance of 0.3, we replace the positive question with a rephrased one. During generation, we prompt GPT-4 to focus on the nouns and actions in the questions, to ensure the generated questions reflect the same video moment with the original question. We show in Fig. 7 some generated examples.

Others. We train all models 10∼20 epochs with initial learning rate of 1e-5. Earlier stopping is adopted if the validation results do not increase in 5 epochs. The batch size is set to 64 for dual-style models and 4∼6 for stacked-style ones. During inference, to fuse the temporal windows derived from Gaussian masking and temporal attention, we simply choose the overlap area of two windows as the final prediction. If there is no overlap, we choose the predictions from temporal attention for better performance.

Table 5. Results of using different number of Gaussian masks.

Model	#Masks	Acc@QA	Acc@GQA	mIoP	mIoU
Temp[CLIP] (NG)	1	59.4	15.5	25.8	7.7
	3	57.9	15.2	25.6	9.1
	5	58.8	15.8	25.7	9.1
	7	58.2	15.4	25.7	10.9

A.3. Additional Experiments

A.3.1 Can multiple Gaussian masks help?

We take Temp[CLIP] with Naive Gaussian (NG) grounding approach to study the effect of using different number of Gaussian masks. The results in Tab. 5 show that using multiple Gaussian masks will hurt the QA accuracy though it increases the grounding performance according to IoU value. The best grounded QA (Acc@GQA) result is achieved by using 5 Gaussian masks. Nonetheless, the improvement over a single Gaussian mask is negligible, *e.g.*, from 15.5% to 15.8%. Therefore, we by default use a single Gaussian mask in major experiments. This also brings higher efficiency.

A.3.2 Does the generated questions help?

We additionally study the effect of extended positive questions in the NG+ method. As shown in Tab. 6, we find that it improves the QA results (Acc@QA) but not for grounded QA (Acc@GQA). In terms of grounding, it brings slightly higher IoU result yet lower IoP compared with the models without using the generated questions. We use the generated positive questions in our final experiments, considering that it improves QA and does not hurt grounded QA. The benefit could be more significant if we rephrase for more questions;

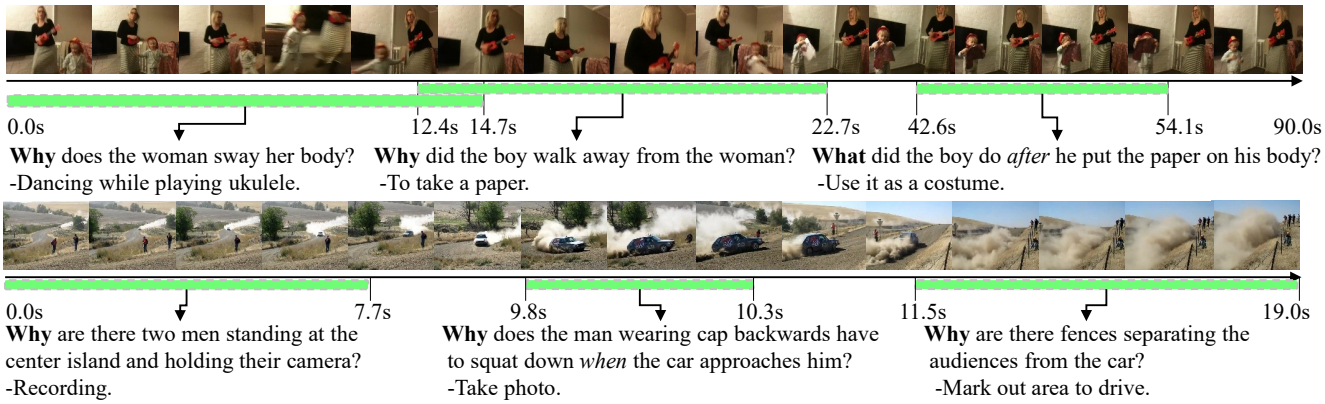


Figure 6. Examples of annotations in NExT-GQA.

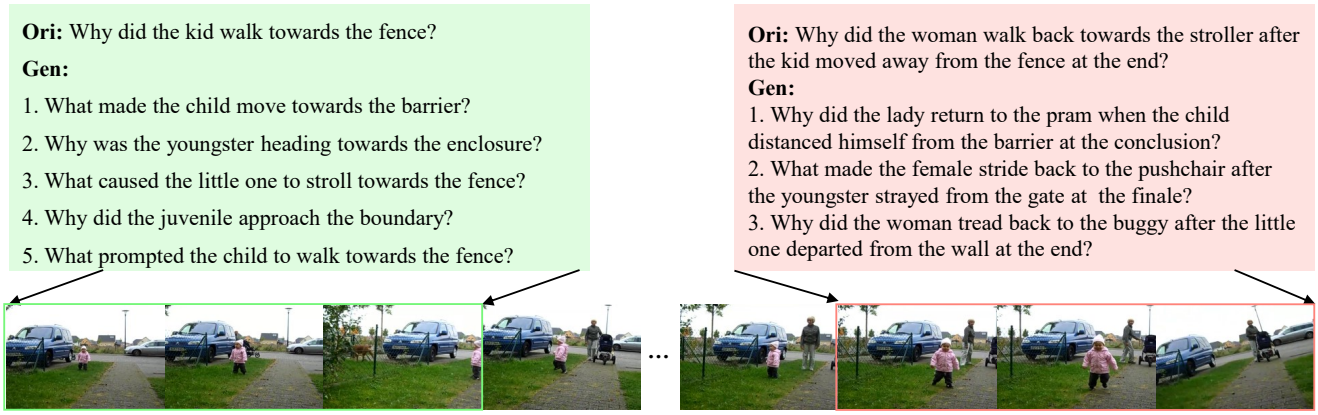


Figure 7. Examples of generated questions by GPT-4.

currently, we only rephrase for 10% of the questions in the training set. Yet, this will result in additional compute cost.

Table 6. Performances without (w/o) using generated questions.

	Model	Acc@QA	Acc@GQA	mIoP	mIoU
NG+	Temp[CLIP] (w/ Gen)	60.2	16.0	25.7	12.1
	Temp[CLIP] (w/o Gen)	59.3	16.0	26.7	9.9
NG+	FrozenBiLM (w/ Gen)	70.8	17.5	24.2	9.6
	FrozenBiLM (w/o Gen)	70.2	17.5	24.4	8.7

A.3.3 Model Efficiency

We discuss the efficiency of Temp[CLIP] and FrozenBiLM in the visually-grounded QA task. For Temp[CLIP], all results are obtained with 1 A5000 GPU. For FrozenBiLM without NG+, the experiment was conducted on 4 A5000 GPUs; for FrozenBiLM with NG+, we run with 4 R8000 GPUs as the model needs about 46G per GPU memory. The time is reported based on 1 epoch over the training and validation data respectively. The results in Tab. 7 show that our grounding module introduces little additional parameters for training and inference compared with the respective backbone models. Yet, the NG+ method takes more time to train. Another observation is that the Temp[CLIP] has much higher training and inference speed than FrozenBiLM.

Table 7. Model Efficiency.

Model	Train Param.	Infer Param.	Model Size	Time (Train)	Time (Infer)
Temp[CLIP]	130.3M	130.3M	0.5G	2.0m	10.0s
w/ NG	130.6M	130.6M	0.5G	2.0m	10.0s
w/ NG+	130.6M	130.6M	0.5G	3.5m	10.0s
FrozenBiLM	29.7M	1.2B	3.8G	0.3h	1.0m
w/ NG	43.9M	1.2B	3.8G	1.3h	1.8m
w/ NG+	43.9M	1.2B	3.8G	3.8h	1.8m

A.3.4 Generalization to Video-LLMs

We study whether our grounding methods (post-hoc, NG and NG+) generalize to more recent multimodal large language models (MLLMs). We take Video-LLaMA [69] as an example. Video-LLaMA takes advantages of frozen LLaMA [48] and pretrains Video Q-Former to bridge video inputs with LLaMA. It has demonstrated good VideoQA performance. To study its performance on NExT-GQA, we outline our adaptation as follows.

First, we omit the audio stream in Video-LLamA as NExT-GQA emphasizes visual grounding. Then, we find that the intermediate video Q-Former cuts off the direct correspondence between video frames/segments and answer outputs. This prevents a post-hoc analysis. To circumvent the Q-Former yet also enjoy its cross-modal pretrained weights, we sample 32 video segments for each video and encode each segment by average-pooling the outputs of Q-Former. The segment representations, versus the original global Q-Former outputs, , are fed to LLaMA along with the QA texts (following the format in LLaMA-VQA [20]) for answer decoding. Moreover, we summarize the Top-K ($K = 6$ is the maximal answer length) prediction scores of each video token as its confidence score for post-hoc temporal analysis. Besides, we prepend a special token to the video token sequence to predict the Gaussian parameters. For NG+, the large model size prevents joint training the two terms (Eqn.3 of the main paper) on our server. As a remedy, we apply a two-stage paradigm by first training for question *grounding* and then fine-tuning for *grounded QA*. Finally, to study the effect of multimodal *video* pretraining, we include a model variant by substituting the Q-Former representations of the segments with CLIP features of their middle frames.

Tab. 8 highlights the following observations of Video-LLaMA' behavior on NExT-GQA: **1)** NG and NG+ give consistent improvements over a post-hoc method. **2)** Like our existing findings in the main paper, there is a large gap between QA and GQA accuracy. **3)** Pretrained Video Q-Former improves over image-text pre-trained CLIP for QA but not video grounding. Tab. 9 gives a comparison between Video-LLaMA and the two major backbones (TempCLIP and FrozenBiLM) in the main paper. We find that Video-LLaMA indeed shows higher Grounded QA (GQA) performance than non-LLM method Temp[CLIP]. However, like FrozenBiLM, the higher GQA accuracy of Video-LLaMA is resulted from its strong QA performance but not because of better grounding. In addition, we find that Video-LLaMA generally performs worse than FrozenBiLM in this task. We believe this is because Video-LLaMA solves QA by exploiting the LLMs to generate the answer word by word, while FrozenBiLM directly classifies each candidate as correct or incorrect answer which is more tailored-made for multi-choice QA. Similar findings can be found in the FrozenBiLM [62] paper which emphasizes the superiority of bi-directional pretrained LLMs to generatively trained ones for classification-based VideoQA.

A.3.5 Result Visualization

We show some prediction cases in Fig. 8. Both models predict the correct answer with reasonable visual grounding results for Q1 and Q2. From the 3rd question, we show that the models suffer a lot in either correctly answering the questions (*e.g.*, Q5, Q6 FrozenGQA and Q8) or providing the right visual evidence for the correct answers (*e.g.*, Q3 FrozenGQA, Q4 and Q7). From the failure examples, we find that when the visual concepts in the answers present throughout the videos (*e.g.* "`grass`" and "`snow`" in Q4 and Q7 respectively), the models can easily predict the correct answers without the need to truly localizing the questioned video segments. Furthermore, the models are still weak in 1) answering the questions which involve small visual objects and 2) substantiating the answers when the visual evidence only takes small portion of the videos (Q4 $\sim$ Q8).

Table 8. Results on NExT-GQA validation set. †: Full validation set of NExT-QA. *: 2-stage training.

Backbone	Method	Acc@QA	Acc@QA†	Acc@GQA	mIoP	mIoU
Video-LLaMA(7B) (CLIP-VIT)	Post-hoc	63.3	65.1	15.6	23.0	8.3
	NG	64.3	67.2	16.5	24.9	**11.4**
	*NG+	**66.7**	**69.8**	**17.2**	**25.2**	10.5
	Improves	+3.4	+4.7	+1.6	+2.2	+2.2
Video-LLaMA(7B) (VQ-Former)	Post-hoc	66.0	68.4	15.5	21.2	5.3
	NG	66.9	69.4	**18.2**	**25.1**	**7.3**
	*NG+	**68.5**	**71.4**	17.4	24.1	6.8
	Improves	+2.5	+3.0	+1.9	+2.9	+1.5

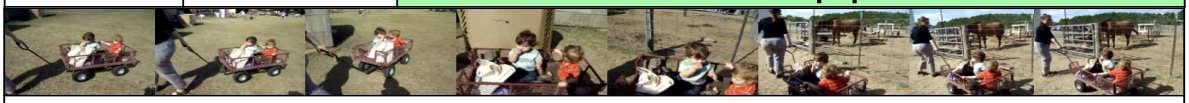
Table 9. Comparison on NExT-GQA test set

Method	Backbone	Acc@QA	Acc@QA†	Acc@GQA	mIoP	mIoU
NG	TempCLIP(130M)	59.4	62.7	15.5	**25.8**	7.7
	Video-LLaMA(7B)	65.1	68.3	16.6	24.9	7.7
	FrozenBiLM(1B)	**70.4**	**73.1**	**17.2**	24.0	**9.2**
NG+	TempCLIP(130M)	60.2	63.3	16.0	**25.7**	**12.1**
	Video-LLaMA(7B)	67.3	70.6	17.1	24.5	11.0
	FrozenBiLM(1B)	**70.8**	**73.1**	**17.5**	24.2	9.6

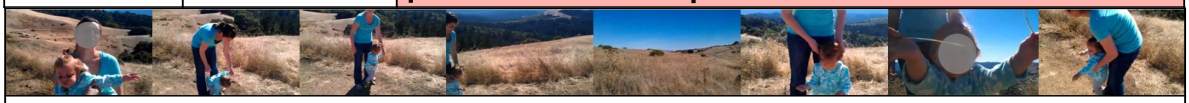
A.4. Discussion on Multi-Choice QA

Popular open-ended VideoQA datasets, such as MSRVTT-QA, MSVD-QA and TGIF-QA, consist of very short videos, typically ranging from 3 to 15 seconds. They do not necessitate temporal grounding. While ActivityNet-QA contains long videos, a large portion of its questions are simple and can be answered with a single frame (by human). Given the above consideration, we experiment on NExT-QA, specifically on its multi-choice QA task as there is currently not much literature oriented for open-ended QA. Multi-choice QA tends to be more susceptible to language bias and spurious vision-language correlation. Because the provided negative answers may not always be distractive enough to challenge the selection of the correct answer without video consultation. Also, the visual concepts mentioned in the negative answers may not exist in the given videos at all. Conversely, our defined grounded-QA task would largely prevent or discourage such short-cut learning.

Q1: What does the dog do after the lady in front reach out her hand in the middle?		11.0s
GroundTruth	Climb onto lady	7.6s – 11.0s
TempGQA	Climb onto lady	8.7s – 11.0s
FrozenGQA	Climb onto lady	7.3s – 8.2s



Q2: How is the horse domesticated and prevented from running away?		26.0s
GroundTruth	Kept within fence	20.0s – 26.0s
TempGQA	Kept within fence	18.2s – 25.7s
FrozenGQA	Kept within fence	19.2s – 22.2s
Q3: Why did the boys turn to their right when they went past the cages?		
GroundTruth	Look at horse	21.5s – 26.0s
TempGQA	Look at horse	18.2s – 25.7s
FrozenGQA	Look at horse	1.7s – 10.2s



Q4: Why does the baby put out her hand near the end?		69.0s
GroundTruth	Show grass	48.8s – 59.2s
TempGQA	Show grass	9.5s – 10.2s
FrozenGQA	Show grass	6.2s – 6.2s
Q5: What does baby do after getting on the ground?		
GroundTruth	Pick up something	8.3s – 11.4s
TempGQA	Drinking milk.	10.1s – 10.2s
FrozenGQA	Drinking milk.	6.2s – 6.2s
Q6: Why does the lady bent down after putting the baby on the ground?		
GroundTruth	Support the baby	2.9s – 11.3s
TempGQA	Support the baby	6.2s – 10.2s
FrozenGQA	Sit down.	29.2s – 29.2s

Q7: What does the man in red shirt do as the man blue slides pass him in the middle?		15.0s
GroundTruth	Throw snowball	6.0s – 7.3s
TempGQA	Throw snowball	7.7s – 10.7s
FrozenGQA	Throw snowball	0.7s – 2.2s
Q8: Why did the man in red shirt sit in the middle of the slope?		
GroundTruth	Look at horse	6.0s – 7.3s
TempGQA	Wait for people	7.7s – 10.7s
FrozenGQA	Wait for people	0.7s – 1.7s

Figure 8. Result visualization on NExT-GQA. TempGQA and FrozenGQA denote Temp[CLIP] and FrozenBiLM with our NG+ grounding mechanism. The ground-truth and correct predictions are in green, while the wrong predictions are in red.