Can I Trust Your Answer? Visually Grounded Video Question Answering

Junbin Xiao, Angela Yao*, Yicong Li, Tat Seng Chua
National University of Singapore

Highlight

CVPR SEATTLE, WA JUNE 17-21, 2024

Abstract

To what extent are the answers of vision-language-models (VLMs) truly anchored at the visual content, and more precisely the "relevant part" of the visual content? In this paper, we have made a thorough analysis by proposing visually-grounded VideoQA task. **Grounded VideoQA requires the models to answer the questions and simultaneously localize the relevant video moments to substantiate their predictions**. To facilitate the study, we construct NExT-GQA, a dataset by extending NExT-QA ' QAs with temporal location labels (see Fig.2). We then analyze a series of VLMs that have reported SoTA results on NExT-QA. To our surprise, **we find that all models are struggling with answer grounding despite their strong QA performance (16% *vs.* 69%)**. As a remedy, we propose a Gaussian-based answer grounding method that learns a Gaussian mask along the temporal dimension of the video for each QA pair, via both QA and VQ contrastive learning. Our experiments demonstrate that this method effectively improves grounding and grounded QA. Yet the gap with human performance is still large (17.5% *vs.* 82.1%), which underscores the need on continued research efforts. Our dataset and code are available at https://github.com/doc-doc/NExT-GQA.

Introduction

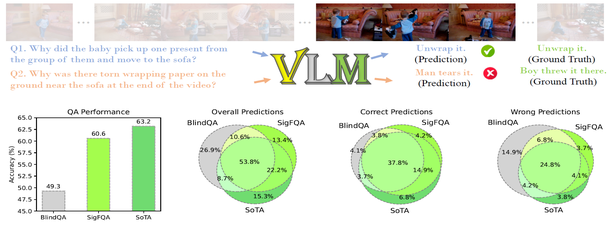

Figure 1. Analysis of predictions on NExT-QA.

- VLM' predictions largely overlap that of its language counterpart and that with the input of a random video frame.
- The models may not learn from causal visual content but more likely from language short-cut and irrelevant visual context.

Contribution

- We conduct the first study of weakly-grounded VideoQA, and release the NExT-GQA benchmark to facilitate research on more trustworthy VLMs.
- We comprehensively analyze a wide range of advances VLMs and reveal their limitation in performing visually grounded QA.
- We propose a simple yet effective Gaussian mask learning method that can be integrated into exiting VLMs to enhance their capabilities of answer grounding and grounded question answering.

NExT-GQA Dataset

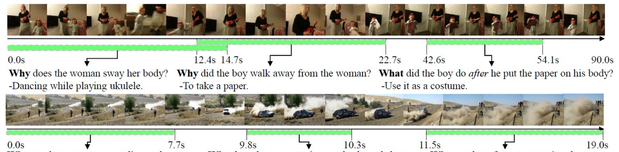

Figure 2. Examples of NExT-GQA.

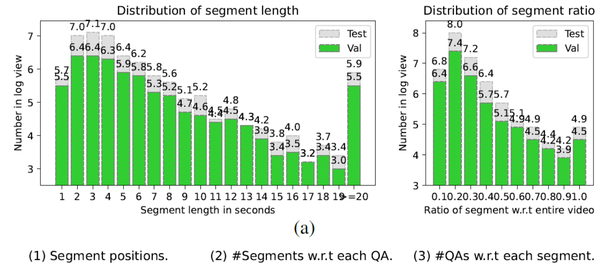

Figure 3. Analysis of temporal labels.

Method to Achieve Weakly-grounded VideoQA

- **Post-hoc Attention:**
  Learn the temporal attention regarding the video tokens.
- **Naïve Gaussian (NG):**

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

$$F_{v_t} = t \cdot \operatorname{softmax}\left(\frac{F^K (F^Q)^\top}{\sqrt{d_k}}\right) F^V$$

$$t \sim N(\mu, \sigma^2)$$

$$t = (\mu - \gamma\sigma, \mu + \gamma\sigma) * d$$

- **Naïve Gaussian with VQ Correspondence Learning (NG+):**

Method to Achieve Weakly-grounded VideoQA

Figure 4. Illustration of our Gaussian mask learning for grounded VideoQA.

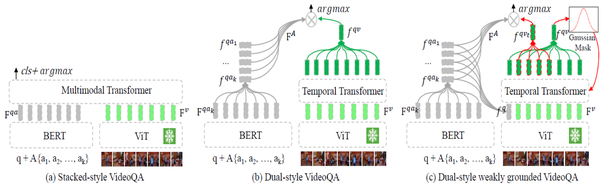

Figure 5. Illustration of integrating Gaussian Mask learning into Transformer architectures.

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

Pull v_t close to the query question, but far from other questions.

Experiment

Q1: To what extent are the answers visually-grounded?

	Model	D/S	CM	Vision	Text	Acc@QA	Acc@QA†	Acc@GQA	mIoP	IoP@0.5	IoP@0.3	mIoU	IoU@0.5	IoU@0.3
	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	Res-Net	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	N	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	6.6	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 1. Results on NExT-GQA test set.

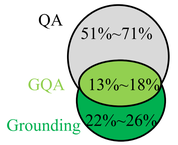

Figure 6. Extent of grounded QA.

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

Table 2. Our method effectively improves 3 different backbones.

Experiment

Q2: Does better QA imply better grounding and vice versa?

- Better QA is not necessarily established by better grounding.
- Having grounding is better than no-grounding, yet correct grounding does not promise correct QA.

	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

Table 3. QA results under different settings.

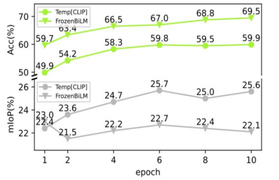

Figure 7. VQA and VG results w.r.t. training epochs.

Q3: How effective is our Gaussian Mask method?

- NG and NG+ consistently boost grounded QA performance (Table 1, 2 & 4), especially in answering questions that truly need video understanding and temporal grounding (Table 5).

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 4. Results with Video-LLaMA.

Models	QA Set	Acc@QA	Acc@GQA	mIoP	mIoU
TempCLIP	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

Table 5. Results on different subset.

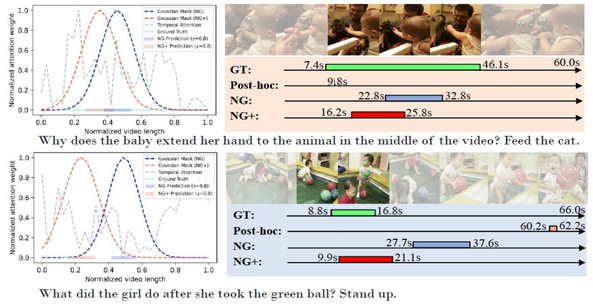

Figure 8. Visualization of the prediction examples.

Conclusion

1. 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.
2. Temporal localization of questions is still a difficult and open challenge. Our studies indicate that solving this problem can largely benefit trustworthy VQA.
3. Our Gaussian Mask method effectively improves grounded QA. However, the great gap with human performance calls for a serious amount of continued research.
4. We hope our NExT-GQA benchmark can contribute towards advancement in these areas.