

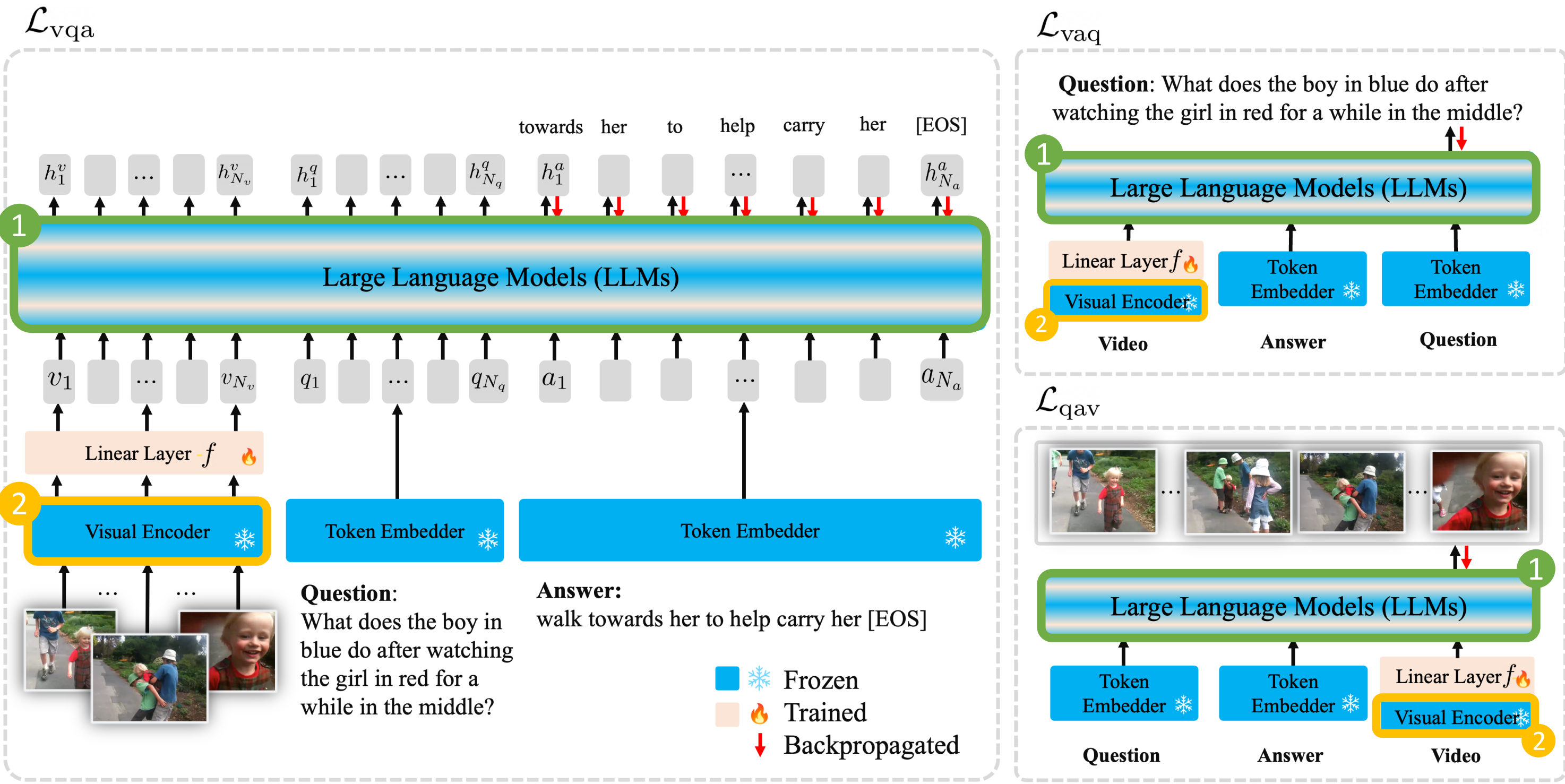
CatchingSTAR

Group4
吳韋論 R12922075
吳浩平 R12922068
江子涵 R12922082
涂子峻 R12942099

Abstract

Based on the Flipped-VQA architecture, which uses three types of loss to train the model for predicting A, Q, and V given VQ, VA, and QA pairs, we introduce several novel and technically significant contributions. Firstly, we replace the visual encoder with ViCLIP, a simple video CLIP designed for transferrable video-text representation. Additionally, we upgrade from LLAMA1-7B to LLAMA2-7B, a more powerful Large Language Model. These alterations are expected to increase the model's performance, enabling better visualization of option probabilities and video frames corresponding to each question.

Architecture



1 LLAMA2

- Foundation language models, based on the Transformer architecture, employing RMSNorm, SwiGLU activation, and Rotary Positional Embedding
- Trained on 45TB, 2 trillion tokens, and the pre-training context length is 4096


2 ViCLIP

- Align video-text representation, initialized from pretrained CLIP
- Integrates spatiotemporal attention in lieu of ViT's self-attention
- Trained on InternVid dataset (7M videos, 234M clips with captions) and incorporates both constrastive learning and mask modeling techniques

Visualization

Interaction


Question:
Which object was opened by the person?



Choises:
a. The laptop (0.04)
b. The refrigerator (0.64)
c. The bag (0.16)
d. The closet/cabinet (0.15)

Sequence

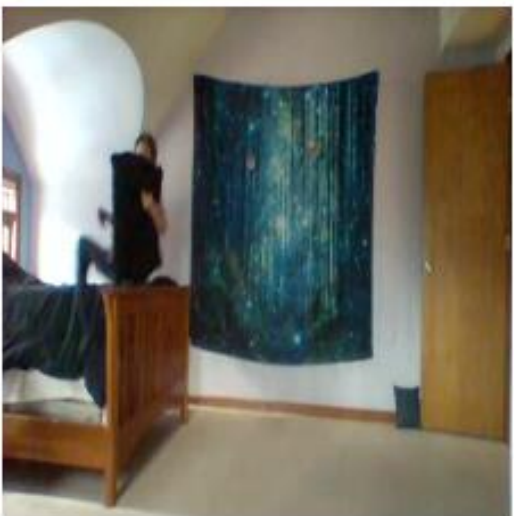
Question:
Which object did the person throw after they closed the door?



Choises:
a. The towel (0.18)
b. The shoe (0.62)
c. The clothes (0.13)
d. The broom (0.07)

Prediction


Question:
What will the person do next with the bed?



Choises:
a. Eat (0.03)
b. Take (0.06)
c. Lie on (0.54)
d. Sit on (0.37)

Feasibility

Question:
What is the person able to do when they are on the side of the shelf?



Choises:
a. Close the closet/cabinet (0.37)
b. Take the shoe (0.16)
c. Hold the laptop (0.28)
d. Put down the cup/glass/bottle (0.19)

We conducted experiments for various settings, tried different LLAMA versions, and tested the differences in performance with various loss configurations. Also, we implemented experiments for replacing the video encoder and for voting mechanisms.

Experiments

Method	Loss	Int_Acc (↑)	Seq_Acc (↑)	Pre_Acc (↑)	Fea_Acc (↑)	Mean (↑)
LLAMA1 7B	VQA					
	VQA+VAQ					
	VQA+VAQ+QAV					
LLAMA2 7B	VQA					
	VQA+VAQ					
	VQA+VAQ+QAV					
LLAMA2 7B + ViCLIP	VQA+VAQ					
Voting (LLAMA1+2)	VQA+VAQ+QAV					