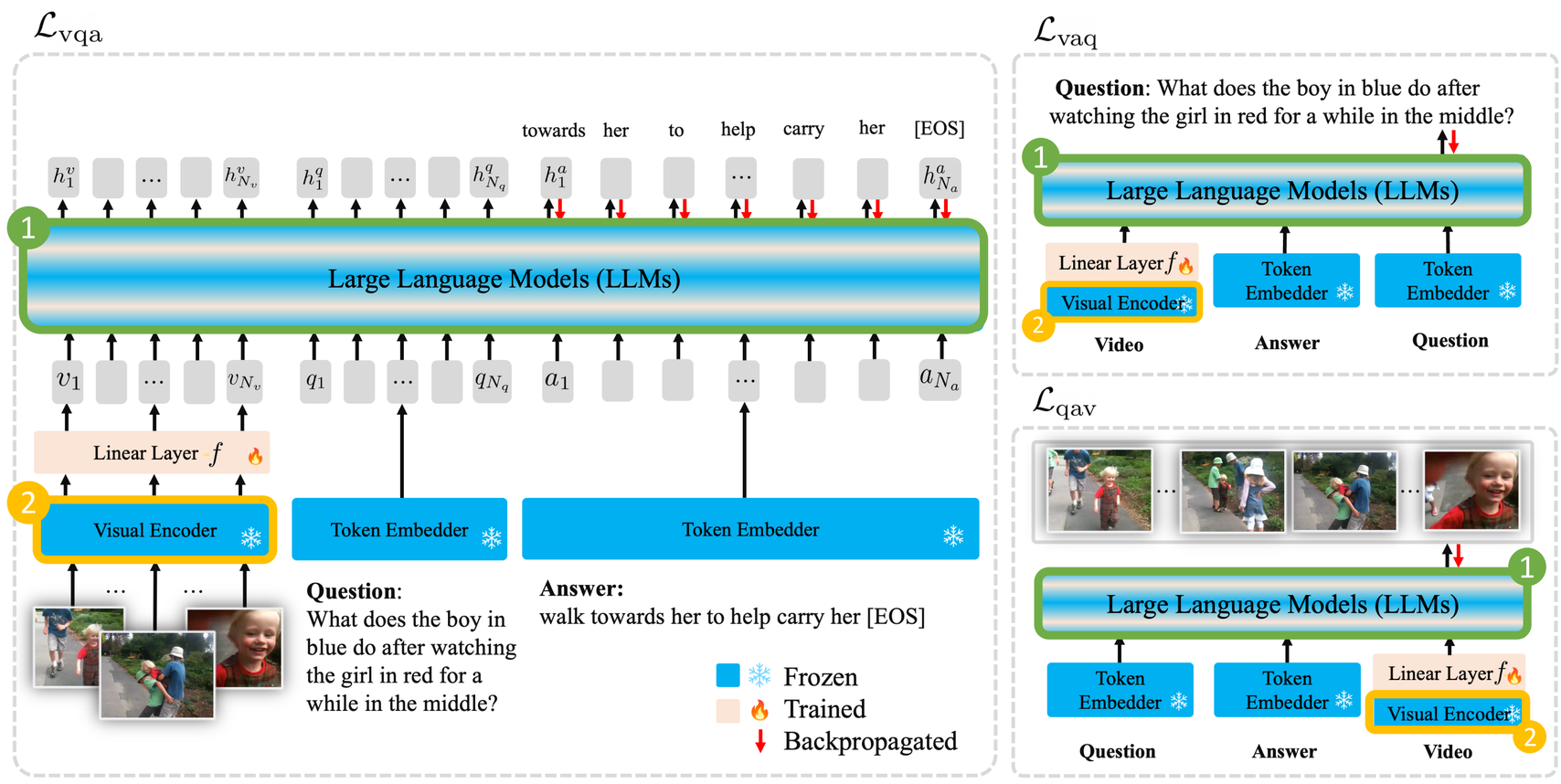
CatchingSTAR

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



LLAMA2

1

- 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

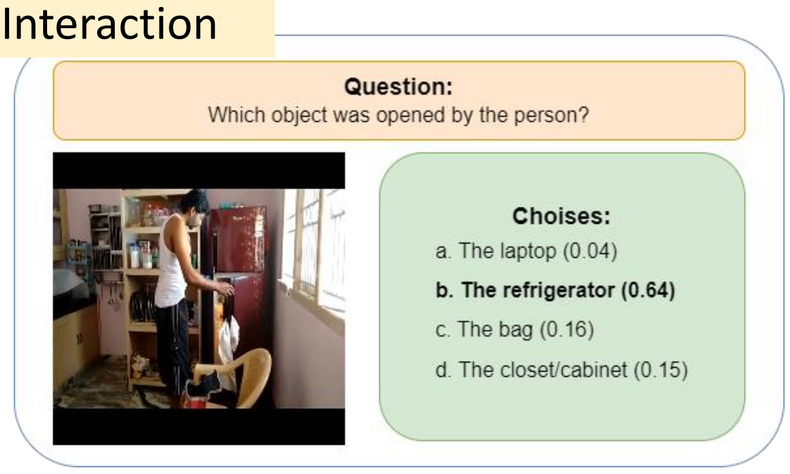
ViCLIP

2

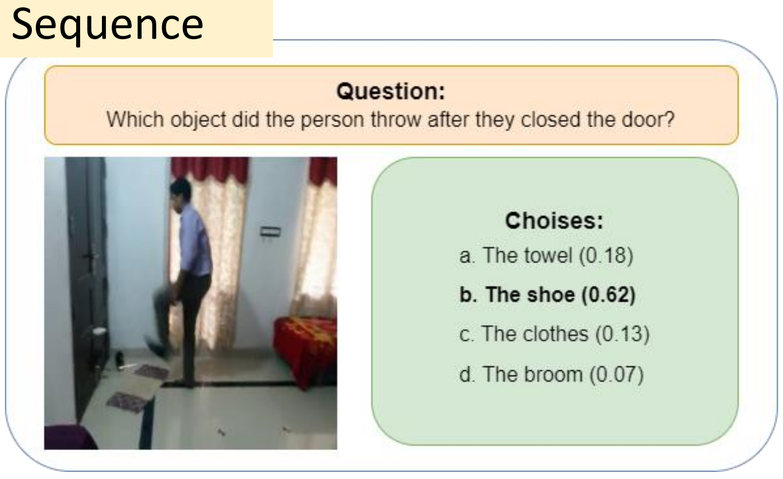
- Align video-text representation, initialized from pretrained CLIP
- Integrates spatiotemporal attention in lieu of ViT's self-attention
- Trained on InternVid dataset (7M videos, 234M clips with captions) and incorporates both constrastive learning and mask modeling techniques

Visualization

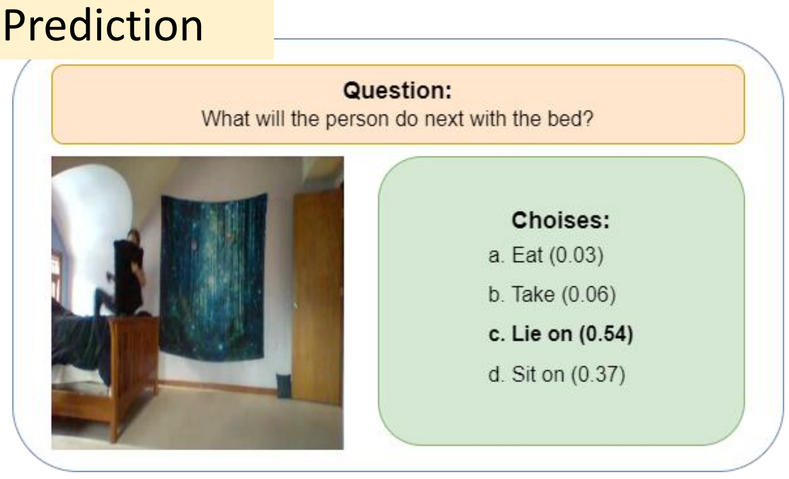
Interaction



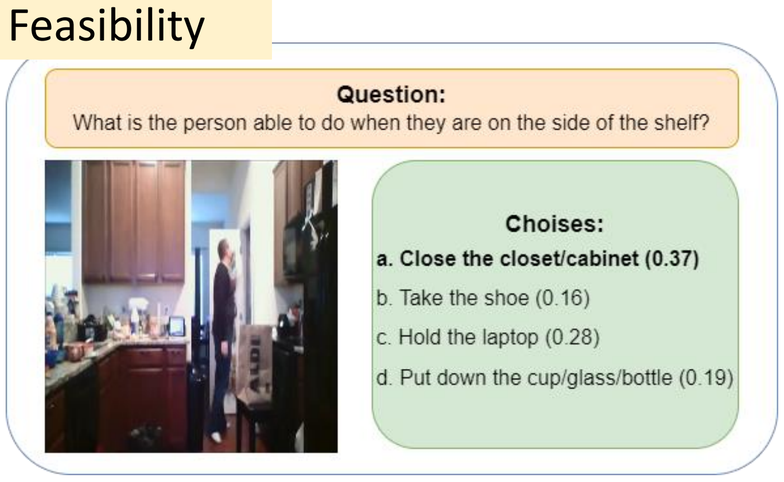
Sequence



Prediction



Feasibility



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					