Modeling Videos: Language as a Key Driver

Research Seminar: Video-Language Models[1]

Lei Wang

Australian National University

December 5, 2023

[1] The works presented in this seminar are sourced from published papers.

A multi-branch cross-modal pre-training framework

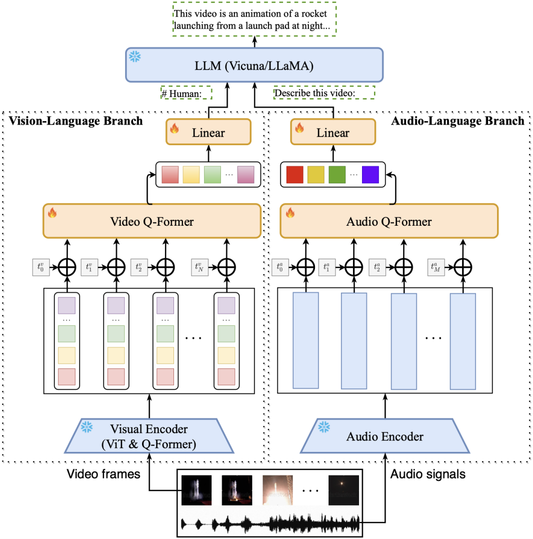


Figure 1: Video-LLaMA.

Prompting visual-language models

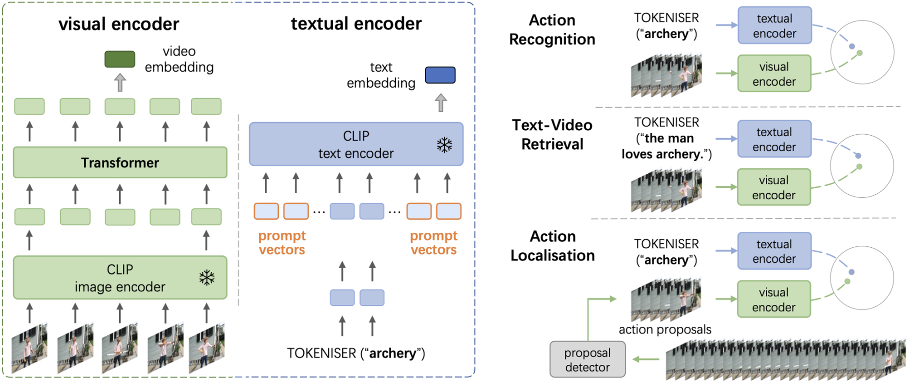


Figure 2: Model adaptation by learning prompts and temporal modelling.

Compositional prompt learning

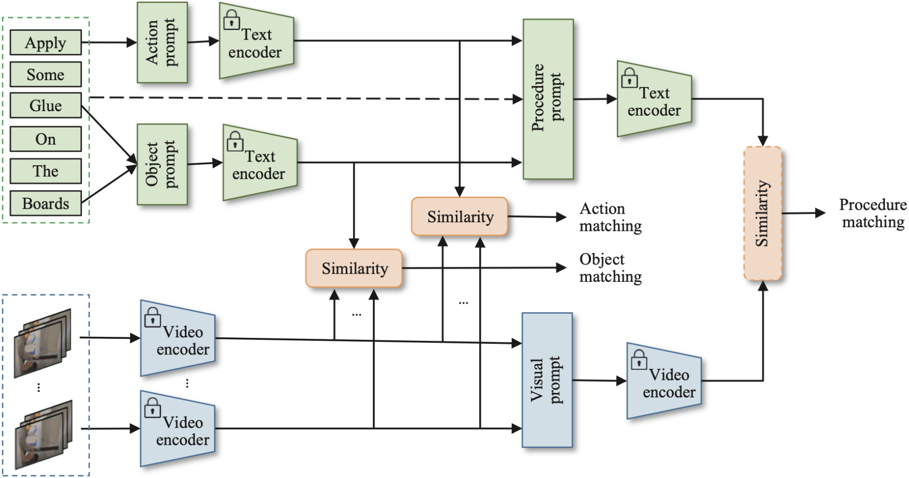
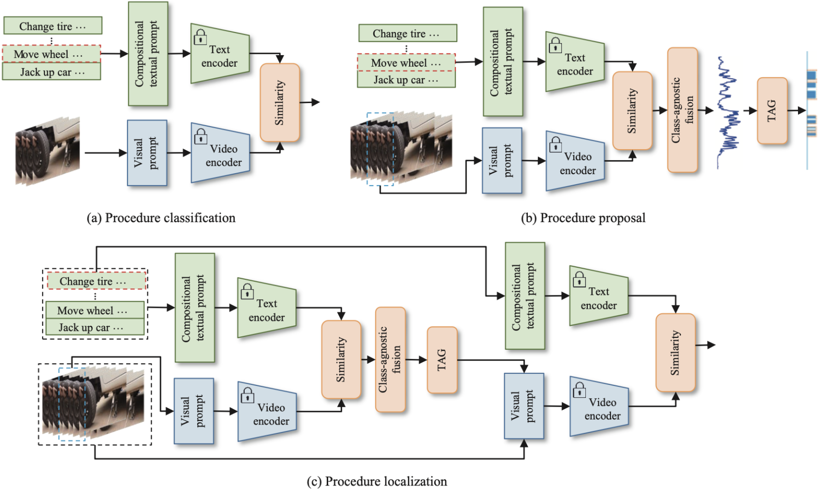


Figure 3: Compositional prompting video-language model.

Compositional prompt learning (cont.)



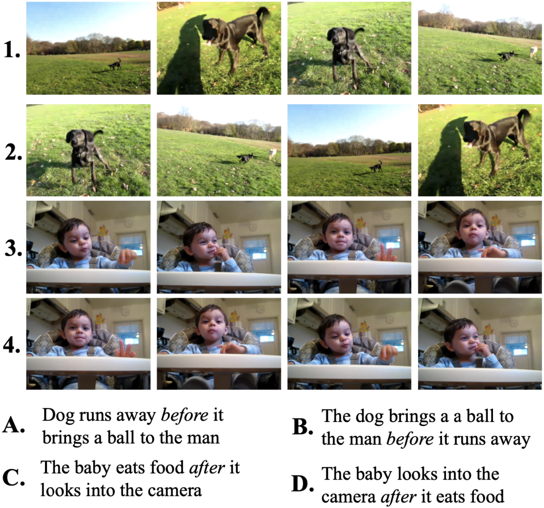
(a) Procedure classification

(b) Procedure proposal

(c) Procedure localization

Figure 4: Three classical downstream tasks for procedure understanding are reformulated as general matching problem.

Instilling video-language models with a sense of time



A. Dog runs away *before* it brings a ball to the man

B. The dog brings a a ball to the man *before* it runs away

C. The baby eats food *after* it looks into the camera

D. The baby looks into the camera *after* it eats food

Figure 5: Understanding the time order of events across video and language is necessary.

Instilling video-language models with a sense of time(cont.)

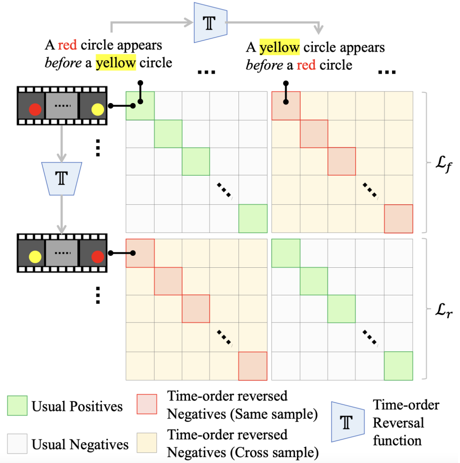
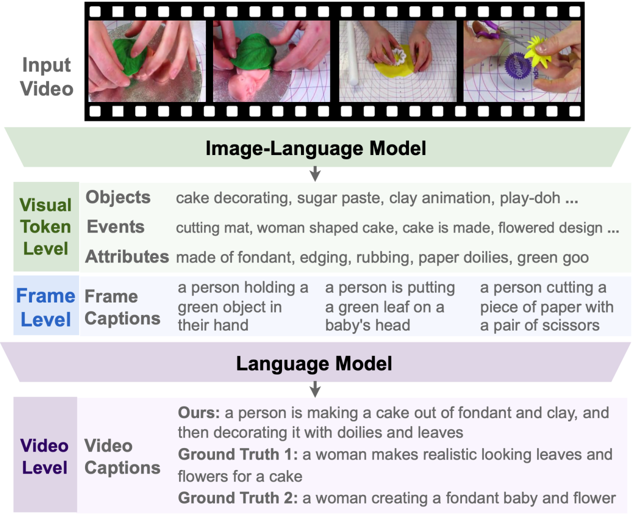


Figure 6: Negatives come from (i) other samples in the batch (cross sample) (ii) time-order reversal within the same sample.

Strong video-language learners w/o pretraining/finetuning



Input Video

Image-Language Model

Visual Token Level

Objects: cake decorating, sugar paste, clay animation, play-doh ...

Events: cutting mat, woman shaped cake, cake is made, flowered design ...

Attributes: made of fondant, edging, rubbing, paper doilies, green goo

Frame Level

Frame Captions: a person holding a green object in their hand | a person is putting a green leaf on a baby's head | a person cutting a piece of paper with a pair of scissors

Language Model

Video Level

Video Captions:

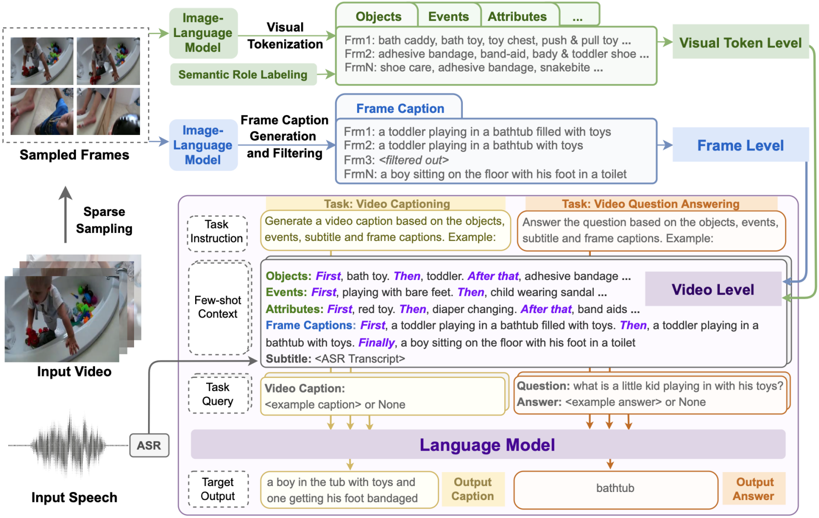
Ours: a person is making a cake out of fondant and clay, and then decorating it with doilies and leaves

Ground Truth 1: a woman makes realistic looking leaves and flowers for a cake

Ground Truth 2: a woman creating a fondant baby and flower

Figure 7: Multiple levels of information in videos.

Strong video-language learners (cont.)



Sampled Frames

Sparse Sampling

Input Video

Input Speech

ASR

Image-Language Model

Visual Tokenization

Semantic Role Labeling

Objects | Events | Attributes | ...

Frm1: bath caddy, bath toy, toy chest, push & pull toy ...
Frm2: adhesive bandage, band-aid, bady & toddler shoe ...
FrmN: shoe care, adhesive bandage, snakebite ...

Visual Token Level

Image-Language Model

Frame Caption Generation and Filtering

Frame Caption

Frm1: a toddler playing in a bathtub filled with toys
Frm2: a toddler playing in a bathtub with toys
Frm3: *<filtered out>*
FrmN: a boy sitting on the floor with his foot in a toilet

Frame Level

Task Instruction

Task: Video Captioning

Generate a video caption based on the objects, events, subtitle and frame captions. Example:

Task: Video Question Answering

Answer the question based on the objects, events, subtitle and frame captions. Example:

Few-shot Context

Objects: ***First***, bath toy. ***Then***, toddler. ***After that***, adhesive bandage ...
Events: ***First***, playing with bare feet. ***Then***, child wearing sandal ...
Attributes: ***First***, red toy. ***Then***, diaper changing. ***After that***, band aids ...
Frame Captions: ***First***, a toddler playing in a bathtub filled with toys. ***Then***, a toddler playing in a bathtub with toys. ***Finally***, a boy sitting on the floor with his foot in a toilet
Subtitle: <ASR Transcript>

Video Level

Task Query

Video Caption:
<example caption> or None

Question: what is a little kid playing in with his toys?
Answer: <example answer> or None

Language Model

Target Output

a boy in the tub with toys and one getting his foot bandaged

Output Caption

bathtub

Output Answer

Temporal-Aware Few-shot Prompt

Figure 8: Representing a video in a unified textural representation containing 3 semantic levels.