DANTE-AD: Dual-Vision Attention Network for Long-Term Audio Description

Adrienne Deganutti, Dr Simon Hadfield, Dr Andrew Gilbert

University of Surrey

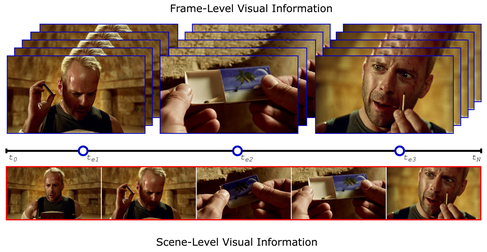
Read the paper

Motivation

- Globally, an estimated 43 million people are fully blind.
- Due to the expensive & time-consuming nature of human-generated audio description (AD), its availability is limited to:
  1. The most popular film & TV.
  2. New or recent film & TV.
  3. High budget productions.
- Automated AD enables more consistent and widespread accessibility in media.
- Growing legal requirements are increasing its commercial demand.
- Audio Description is **NOT** video captioning.

	Video Captioning	Audio Description
Primary purpose	Generates a textual description of visual content	Provides a spoken narration of key visual elements for accessibility
Target audience	Machine learning applications, indexing/search systems	Blind or visually impaired audiences
Content focus	Visual events, objects, actions	Visual events, objects, actions, scenes, displayed text, mood
Typical length	Concise 1-2 sentences	Variable, timed with gaps in dialogue

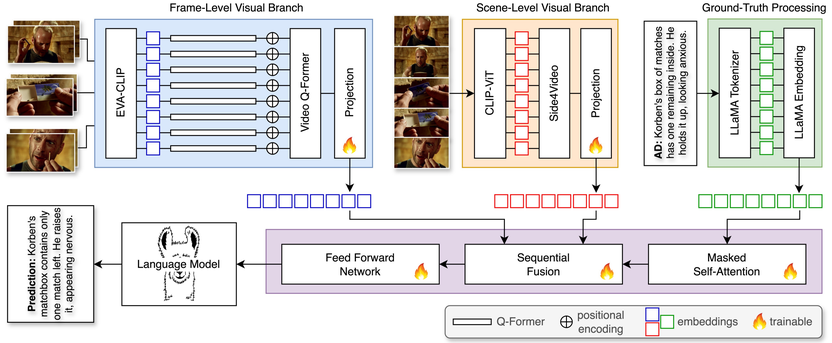
- Our method improves audio description generation by focusing on **contextual awareness** within a movie scene.



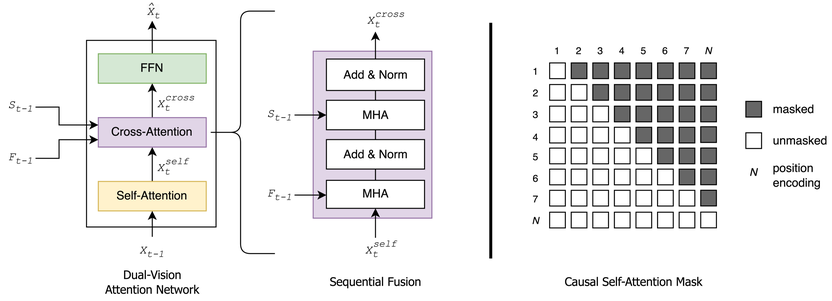
- Frame-level visual embeddings captures detailed object-centric information.
- Scene-level visual embeddings captures global scene information.

Method

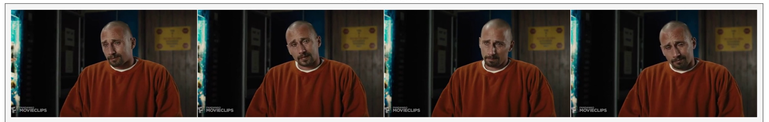
The fusion of frame-level and scene-level visual embeddings improves context awareness for audio description generation.



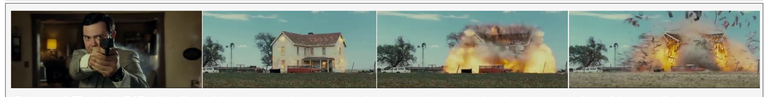
- We extract frame-level visual embeddings using EVA-CLIP and scene-level visual embeddings using Side4Video.
- The dual-vision transformer cross-attends to frame and scene embeddings via a sequential fusion strategy.
- The sequential fusion module first processes frame embeddings to capture fine-grained temporal details. The second layer attends to the scene embeddings to incorporate high-level contextual cues.



Results



Ground Truth: His face contorts with emotion, and he swallows as tears well in his eyes.
DANTE-AD: His eyes fill with tears as he stares down at the floor.
AutoAD III: He looks at the floor.



Ground Truth: O'Reilly fires his gun and an explosion rips through the house sending Moses flying.
DANTE-AD: The explosion blows the roof off the house and sends debris flying.
AutoAD III: The house explodes.

Ground Truth: She stares at him.
DANTE-AD: She stares at him with a worried expression.
AutoAD III: Max looks at her.

Ground Truth: Hubbard is dialing a number.
DANTE-AD: He picks up the phone and dials a number.
AutoAD III: He picks up the receiver.

Method	VLM	LLM	Cr ↑	R@1/5↑	LLM-AD-Eval (%)↑ LLaMA	GPT-3.5
Video-BLIP2	EVA-CLIP	OPT-2.7B	4.8	22.0	31.50	23.33
Video-Llama2	EVA-CLIP	LLaMA2-7B	5.2	23.6	31.83	23.83
AutoAD-II	CLIP-B32	GPT-2	13.5	26.1	34.66	25.50
AutoAD-III	EVA-CLIP	OPT-2.7B	22.3	29.8	46.33	37.50
AutoAD-III	EVA-CLIP	LLaMA2-7B	25.0	31.2	48.67	**38.17**
DistinctAD	$CLIP_{AD}$-B16	LLaMA3-8B	22.7	**33.0**	48.00	-
DANTE-AD	EVA-CLIP + S4V	LLaMA2-7B	**28.89**	28.01	**48.83**	34.50