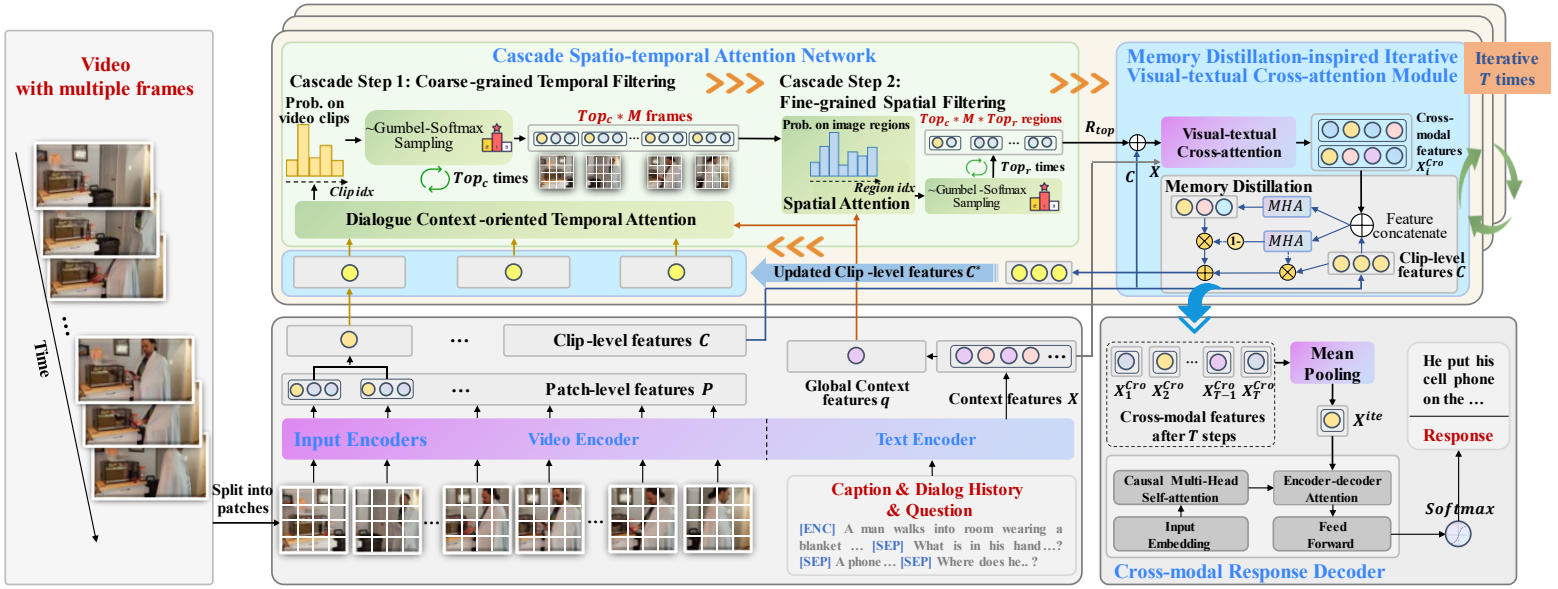
Cascade Context-oriented Spatio-temporal Attention Network for Efficient and Fine-grained Video-grounded Dialogues

Hao WANG, Bin GUO, Mengqi CHEN, Qiuyun ZHANG, Yasan DING, Ying ZHANG, Zhiwen YU

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40387-w

Problems & Ideas

- Problems of conventional video-grounded dialogues:
  - Suffering from identifying context-relevant video parts while disregarding the impact of redundant information in long-form and content-dynamic videos.
  - Existing video-grounded dialogues neglects the sophisticated correspondence between various granularities of visual and textual concepts (e.g., still objects with nouns, dynamic events with verbs).
  - Ideas: Filtering of video information unrelated to dialogue context, iterative semantic information alignment and inference.



The overall architecture of our proposed COSTA, which is composed of (1) the input encoders, consisting of a video encoder and a text encoder, (2) a cascade spatio-temporal attention network with two cascade steps (coarse-grained temporal filtering and fine-grained spatial filtering), (3) a memory distillation-inspired iterative visual-textual cross-attention module, and (4) a cross-modal response decoder

Main Contributions

- Contributions:
  - A cascade attention network to localize only the most relevant video clips and regions in a coarse-to-fine manner which effectively filters the irrelevant visual semantics;
  - A memory distillation-inspired iterative visual-textual cross-attention strategy to progressively integrate visual semantics with dialogue contexts across varying granularities.

Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	BERT-Score
AVSD@DSTC7								
SCGA	0.745	0.622	0.517	0.430	0.285	0.578	1.201	0.515
VGMNM	-	-	-	0.429	0.278	0.578	1.188	0.530
PDC	0.747	0.616	0.512	0.429	0.282	0.579	1.194	0.539
BiST	0.755	0.619	0.510	0.429	0.284	0.581	1.192	0.552
PDC-GPT2	0.770	0.653	0.539	0.449	0.292	0.606	1.295	0.579
RLM	0.765	0.643	0.543	0.459	0.294	0.606	1.308	0.583
CRMSG	0.776	0.652	0.551	0.466	0.304	0.609	1.333	0.588
DialogMCF	0.777	0.653	0.547	0.457	0.306	0.613	1.352	0.592
THAM	0.778	0.654	0.549	0.468	0.308	0.619	1.335	0.595
COSTA	**0.793**	**0.657**	**0.564**	**0.490**	**0.315**	**0.642**	**1.388**	**0.641**
COSTA-BLIP	**0.819**	**0.673**	0.580	**0.508**	**0.325**	0.649	1.410	**0.663**
COSTA-Frozen	0.804	0.664	**0.588**	0.506	0.318	**0.661**	**1.442**	0.651
Video LLaMA	0.780	0.642	0.558	0.483	0.308	0.633	1.373	0.625
w/ COSTA	0.805	0.669	0.570	0.504	0.317	0.660	1.419	0.652
LLaMA Adapter	0.790	0.656	0.562	0.488	0.316	0.641	1.384	0.641
w/ COSTA	0.817	0.674	0.586	0.512	0.327	0.668	1.440	0.666
Video Chat	0.795	0.661	0.565	0.493	0.317	0.644	1.394	0.647
w/ COSTA	0.827	0.680	0.597	0.523	0.327	0.670	1.445	0.686
Video-ChatGPT	0.798	0.667	0.570	0.502	0.322	0.655	1.427	0.661
w/ COSTA	0.840	0.696	0.608	0.541	0.332	0.678	1.553	0.719

Models	Language Fluency	Context Coherence	Factual Correctness	Kappa
AVSD@DSTC7				
RLM	1.62	1.61	1.35	0.65
CRMSG	1.75	1.68	1.52	0.68
THAM	1.78	1.70	1.56	0.69
COSTA	**1.83**	**1.78**	**1.71**	**0.74**
DVD				
HRNN	1.54	1.59	1.28	0.61
VDTN	1.68	1.65	1.45	0.64
MTN	1.72	1.66	1.52	0.67
COSTA	**1.75**	**1.73**	**1.66**	**0.70**

Left: Evaluation results of on the test set of AVSD@DSTC7 of COSTA and baselines; Right: Human evaluation results on the test set AVSD@DSTC7 and DVD of COSTA and baselines.