LoCo-MAD: Long-Range Context-Enhanced Model Towards Plot-Centric Movie Audio Description
Supplementary Material

Jiayi Wang, Zihao Liu, and Xiaoyu Wu ✉

Communication University of China, Beijing 100024, China
{blindwang,liuzihao,wuxiaoyu}@cuc.edu.cn

In this document, we provide additional visualization results and ablation experiments.

A Ablations with AutoAD-II

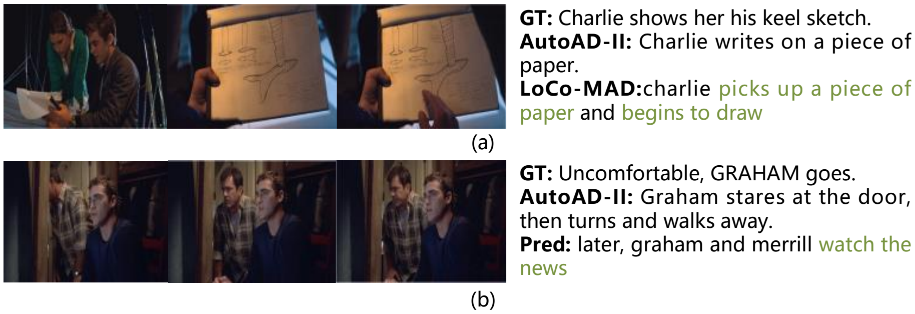
We report the ablations with AutoAD-II [1], which shows our method has noticeable improvement under fair comparison. As shown in Tab. A, our model outperforms AutoAD-II (C 12.2 vs 10.0) without a character recognition module (Char.) and contextual prompts. The Char. proposed in AutoAD-II is trained on character name labels from MovieNet [3], introducing external knowledge beyond the original task. In this way, our model cannot compare with AutoAD-II (row 5 of Tab. A) directly. For models with the contextual prompts, our method significantly outperforms [2] and [5].

Table A: Comparison with AutoAD-II. Chars. and Context refer to the usage of character recognition module and contextual prompts.

Model	Char.	Context	R-L	C	S
AutoAD-II	✘	✘	9.7	10.0	-
LoCo-MAD	✘	✘	**10.9**	**12.2**	**3.3**
AutoAD [2]	✘	✔	11.9	14.3	4.4
MM-Narrator [5]	✘	✔	12.1	11.6	4.5
AutoAD-II	✔	✔	13.4	**19.5**	-
LoCo-MAD	✘	✔	**13.9**	18.8	**5.5**

B Qualitative comparison with AutoAD-II

As shown in Fig. A, we report the qualitative comparison with AutoAD-II on MAD-v2 [2]. We use the predictions of AutoAD-II directly from the original paper. Our model accurately describes character identities and actions, demonstrating comparable performance with AutoAD-II.



GT: Charlie shows her his keel sketch.
AutoAD-II: Charlie writes on a piece of paper.
LoCo-MAD:charlie picks up a piece of paper and begins to draw

GT: Uncomfortable, GRAHAM goes.
AutoAD-II: Graham stares at the door, then turns and walks away.
Pred: later, graham and merrill watch the news

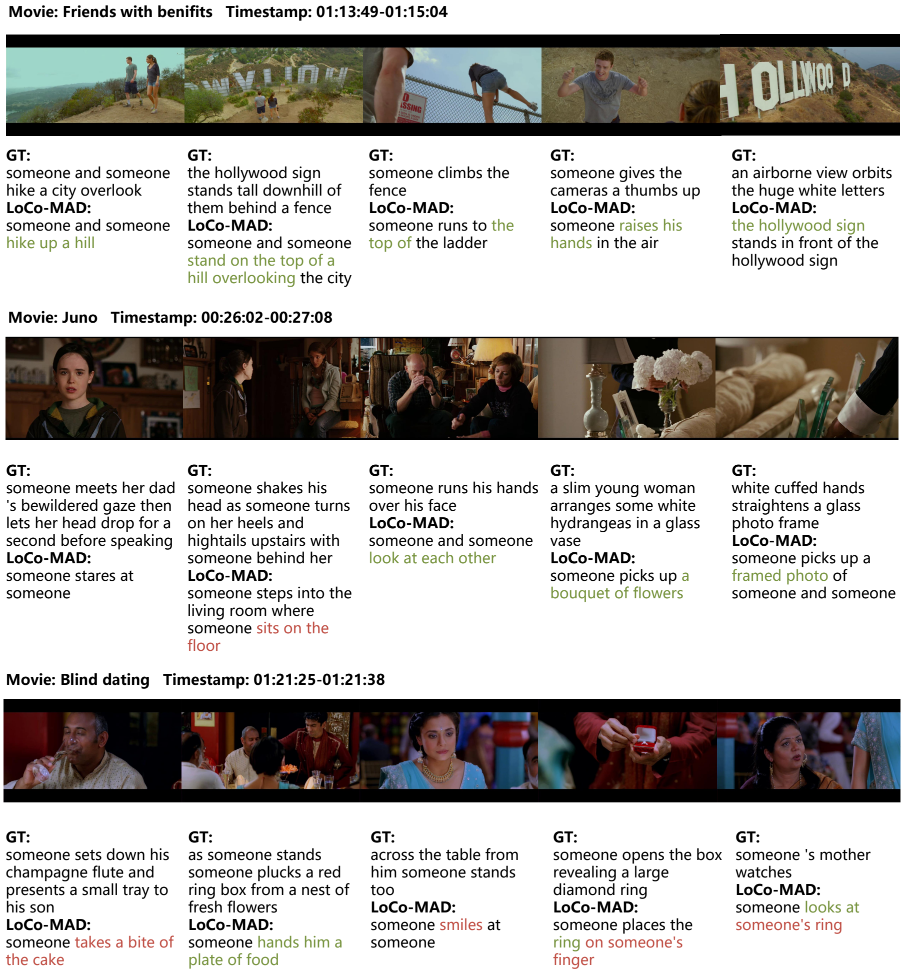
Fig. A: Qualitative comparison with AutoAD-II on MAD-v2 dataset. The movies are from: (a): Signs (2002), (b): Charlie St. Cloud (2010).

C Quality examples on LSMDC dataset

As shown in Fig. B, we report the qualitative examples on LSMDC [4] dataset. Each examples contain five consecutive movie clips. The results are largely accurate in object recognition and character movement recognition (highlighted in green), but there are still some hallucinations (highlighted in red).

References

1. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad II: the sequel - who, when, and what in movie audio description. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 13599–13609. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.01255
2. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad: Movie description in context. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 18930–18940. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01815
3. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Computer Vision - ECCV 2020 - 16th European Conference. Lecture Notes in Computer Science, vol. 12349, pp. 709–727. Springer, Glasgow, UK (2020). https://doi.org/10.1007/978-3-030-58548-8_41
4. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C.J., Larochelle, H., Courville, A.C., Schiele, B.: Movie description. Int. J. Comput. Vis. **123**(1), 94–120 (2017). https://doi.org/10.1007/S11263-016-0987-1
5. Zhang, C., Lin, K.Q., Yang, Z., Wang, J., Li, L., Lin, C.C., Liu, Z., Wang, L.: Mm-narrator: Narrating long-form videos with multimodal in-context learning. ArXiv **abs/2311.17435** (2023)



Movie: Friends with benifits Timestamp: 01:13:49-01:15:04

GT:
someone and someone hike a city overlook
LoCo-MAD:
someone and someone hike up a hill

GT:
the hollywood sign stands tall downhill of them behind a fence
LoCo-MAD:
someone and someone stand on the top of a hill overlooking the city

GT:
someone climbs the fence
LoCo-MAD:
someone runs to the top of the ladder

GT:
someone gives the cameras a thumbs up
LoCo-MAD:
someone raises his hands in the air

GT:
an airborne view orbits the huge white letters
LoCo-MAD:
the hollywood sign stands in front of the hollywood sign

Movie: Juno Timestamp: 00:26:02-00:27:08

GT:
someone meets her dad 's bewildered gaze then lets her head drop for a second before speaking
LoCo-MAD:
someone stares at someone

GT:
someone shakes his head as someone turns on her heels and hightails upstairs with someone behind her
LoCo-MAD:
someone steps into the living room where someone sits on the floor

GT:
someone runs his hands over his face
LoCo-MAD:
someone and someone look at each other

GT:
a slim young woman arranges some white hydrangeas in a glass vase
LoCo-MAD:
someone picks up a bouquet of flowers

GT:
white cuffed hands straightens a glass photo frame
LoCo-MAD:
someone picks up a framed photo of someone and someone

Movie: Blind dating Timestamp: 01:21:25-01:21:38

GT:
someone sets down his champagne flute and presents a small tray to his son
LoCo-MAD:
someone takes a bite of the cake

GT:
as someone stands someone plucks a red ring box from a nest of fresh flowers
LoCo-MAD:
someone hands him a plate of food

GT:
across the table from him someone stands too
LoCo-MAD:
someone smiles at someone

GT:
someone opens the box revealing a large diamond ring
LoCo-MAD:
someone places the ring on someone's finger

GT:
someone 's mother watches
LoCo-MAD:
someone looks at someone's ring

Fig. B: Qualitative examples on LSMDC dataset.