CS 784 Final Project: Midterm Check-in

Eric Huang

March 2, 2025

This is the midterm check-in for the final project. For my word, I have chosen the word 'grounding'.

1 What I've done

1.1 Meta-analysis

Before choosing a word, I decided to first perform some meta-analysis to understand which papers were more important and how overloaded and saturated the term was. To understand which papers were more important, I took the top-p papers (i.e. top papers until we meet some percentage p of the total word count across the entire dataset) based on the relative frequency, i.e. the count of a certain word compared to its total count of words. I performed this per split of the dataset, not mixing up conferences. Note that I simply used the pure substring and not word based on boundaries when calculating this; this ensures that plurality will not cause any miscounts.

After performing this analysis, with $p = 0.1$, I decided to choose the word 'grounding' which had 52 papers total, making it possible to read through all of these. Across all these papers, the highest word frequency was 2.16% found in the ECCV conference for a Llava grounded visual chat [3].

1.2 Paper Reading

I have read a small subset of the top-p papers described above on uncovering the following meanings of the word 'grounding':

1. Image grounding: A fine-grained understanding of images, including specific regions and alignment [3, 1].
2. Video Grounding: Focuses on identifying and localizing specific moments in the video based on descriptions [1].
3. Spatio-Temporal Video Grounding (might just be a very similar task as video grounding): Focuses on identifying a certain object within videos based on query sentences [2].
4. Grounding box: An annotated box utilized within image datasets to provide examples of what the ground truth is. Usually used in instances of trying to identify certain objects [1].

Generally, grounding means developing an understanding of the composition of the input. The term is overloaded in terms of the precise composition based on the modality.

I have also discovered that the terminology of 'understanding' is tied to 'grounding'. Usually, by improving 'grounding', i.e. figuring out how the input works, one can improve the results and reduce hallucinations.

2 What I plan to do next

I need to continue reading through the list of 52 papers I have compiled, and bin them into the different ways that grounding is used. Based on these different meanings, I should rerun some analysis to try and see if I can analyze how many papers use grounding in their different senses.

I should also do some contrastive analysis and look at the bottom-p and middle-p papers to understand whether or not they use grounding within a separate context. Another simple extension is to add in the word grounded, as it seems as though that is another commonly used term in lieu of grounding.

Finally, I plan on building an understanding of how people tackle grounding tasks within multimodal settings alongside common benchmarks used. I will conduct an analysis on the papers that include sufficient 'grounding' instances and analyze the proportion of them that use certain public benchmarks. I plan to explore the benchmark papers

themselves and analyze the citation graph to understand how everything works together.

References

[1] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language enhanced multi-modal grounding model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, Bangkok, Thailand, August 2024. Association for Computational Linguistics. 2

[2] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Video-groundingdino: Towards open-vocabulary spatio-temporal video grounding, 2024. 2

[3] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models, 2023. 1, 2