A Meta-Analysis of "Grounding" in AI/ML Literature

Fatima Sohail
School Of Computer Science
University of Waterloo
f4sohail@uwaterloo.ca

Abstract

In this paper, we conduct a survey targeting the keyword “grounding” to demonstrate different senses of the word. Building our analysis on the AI-paper-crawl dataset from Forty Two AI Lab, we conduct 1) a meta analysis on the dataset to identify the usage and trends for the keyword “grounding” and 2) a literature review to highlight a few common implications of the word.

1 Introduction

The rapid evolution of Artificial Intelligence (AI) and Machine Learning (ML) has led to significant shifts in terminology, with certain terms acquiring multiple meanings across different subfields. One such term, “grounding,” has become particularly ambiguous, encompassing diverse interpretations in areas such as natural language processing, robotics, and cognitive modeling.

Having a dictionary definition of “a knowledge of the basic facts about a particular subject” (Cambridge, 2025), "grounding" is a foundational concept in many AI applications. However, its varied usage has made it difficult to establish a clear, standardized definition.

In this paper, we explore how “grounding” is employed in AI literature, aiming to map out its different meanings and contextual nuances. Using the AI-paper-crawl dataset from Forty Two AI Lab (Seed42Labs), we answer a set of questions about the word and other related words, analyze trends in its usage over time, and examine how different research communities interpret the term. Additionally, we conduct a literature review to identify common themes and implications associated with grounding. By shedding light on these variations, our study seeks to clarify the term’s role in AI discourse.

2 Methodology

The scripts used to create the pipeline for and conduct analysis as per the methodology discussed in this section can be found here: https://github.com/sohailfatima/cs784_project

2.1 Filtering Papers

We start by filtering the entire corpus on the keyword “grounding”. Unsurprisingly, this results in an extremely high number of papers for each conference. We then conduct a TF-IDF search over these papers to extract the 100 most relevant papers (by TF-IDF rank) for the keyword across all conferences. Note that we only use the main paper body for this search to avoid any biases from titles of papers used as references, or from misleading titles or abstract due to limited context. After removing a few duplicates across conferences, this results in 96 unique papers[1] between 2015 and 2024 for a more granular quantitative analysis and literature review.

2.2 Quantitative Analysis

We first utilize the entire dataset obtained in step one of paper filtration to extract some high level insights on temporal trends of papers involving grounding.

We then use the dataset extracted by TF-IDF to conduct finer analysis such as identifying word co-occurrences and sub-field correlations, and clustering papers and topic modeling to uncover similar themes and topics.

We use the resulting data in conjunction with our literature review to reason about trends and evolution of “grounding” in common literature.

[1] List available in the github repository linked above. Note that not all papers have been referenced in our analysis.

2.3 Literature Review

We use the paper abstracts to get an idea of the use of grounding in all 96 papers, and use the paper clusters identified by BERT clustering for more detailed summarization.

2.4 Questions

After summarizing different senses of "grounding", we try to answer the following questions regarding it's usage:

1. How has the presence of "grounding" changed in papers over-time?
2. How does the presence of related keywords change in papers over-time? Which words are more common (and why)?
3. Which ML sub-fields are commonly mentioned in these papers, are there any temporal trends?

3 Common Senses of Grounding

In this section, we will summarize different senses of "grounding" that were prevalent in the papers we filtered. These are largely based on the different contexts in which "grounding" is relevant and what it means in each such context, as opposed to the ways in which ideas and concepts are grounded (e.g. communication, sensorimotor connections) in various domains. Through this, we will highlight the different meanings it conveys.

3.1 Visual Grounding

Visual grounding is the task of associating textual descriptions with corresponding visual elements in images, videos, or 3D data. It plays a crucial role in human-computer interaction, autonomous systems, and multimodal AI and has applications such as image captioning (Wang et al., 2024), phrase localization (Zhu et al., 2022), and vision-language models (Zeng et al., 2024). Models may be grounded via reward modelling (Yan et al., 2024), supervised training (Jiang et al., 2022), or composition-aware finetuning (Zeng et al., 2024). Visual grounding can be further classified into different subtypes based on the type of data and task.

3.1.1 Video Grounding

Video grounding includes application such as video captioning (Jiang et al., 2024; Ma et al., 2020), spatio-temporal secton localization (Wasim et al., 2024) and Video QA (Xiao et al., 2024; Shrestha et al., 2020). This body of work focuses on improving accuracy by linking entities, actions, and language to visual cues across frames.

3.1.2 Visual Grounding and Tracking

A unique application of video grounding in target location for tracking purposes is presented by Zhou et al., 2023. Combining object tracking with natural language-based grounding, it uses a single model to link language descriptions to tracked objects. This research, and similar applications, are a useful step forward for surveillance and autonomous systems using visual grounding.

3.1.3 Multi-Modal Visual Grounding

Multi-modal visual grounding associates or aligns visual content with textual (Shao et al., 2023), linguistic (Guo et al., 2023), or audio/sound based modes (Tian et al., 2021). Such modals learn to jointly ground different modes, which is valuable as it can better mimic real world scenarios. Sound-based techniques in particular are valuable for applications like hearing aids and multimedia retrieval.

3.1.4 Embodied Visual Grounding

A combination of embodied grounding and visual grounding, embodied visual grounding refers to the process by which an AI agent perceives, interprets, and interacts with its environment by linking visual inputs to actionable knowledge through physical experience. It can for example, enable an AI agent to locate objects in a real-world scene using high-level instructions and ground both scenes and objects (Lin et al., 2021). This aspect of visual grounding is especially useful for assistive robotics and autonomous agents.

3.2 Spatio-temporal Grounding

Spatio-temporal grounding also extends visual grounding to video-based tasks, where objects/actions must be located across both space and time. It may be used for action recognition, video question answering, and moment localization (Wasim et al., 2024; Chen et al., 2024; Lee and Sung, 2024; Song et al., 2021; Barrios et al., 2023; Lin et al., 2021; Bao et al., 2021; Ning et al., 2022; Jin and Mu, 2024; Afouras et al., 2023; You et al., 2023; Li et al., 2022; Zhou et al., 2021).

3.3 3D Object and Scene Grounding

Another sub-type of visual grounding, 3D object and scene grounding involves the process of linking textual descriptions directly to 3D environments, allowing machines to associate language with spatial and visual elements in a 3D space (Liu et al., 2021; Sun et al., 2024; Liu et al., 2024). It refers to connecting linguistic information to real-world spatial structures, enabling the identification of objects and events within a 3D context. This could include grounding text to multiple objects in a 3D scene (Zhang et al., 2023) or integrating dense captions with specific points in 3D space (Cai et al., 2022; Chen et al., 2023b; Unal et al., 2024), and uses techniques like view-based grounding (Guo et al., 2023), which allow AI models to use multiple perspectives to better understand and localize objects within the 3D space. This form of grounding is fundamental for robots, autonomous systems, and augmented reality applications, where accurate spatial understanding of objects and environments is essential for effective interaction and decision-making.

3.4 Language Model Grounding

Language model (LM) grounding refers to the process of ensuring factual soundness and alignment with external or human knowledge of the LM outputs. Lee et al. 2023 defines "being grounded" in this context as "(1) fully utilizing the necessary knowledge from the provided context, and (2) staying within the limits of that knowledge", while several other researches define "grounding" as just having correct responses that match human responses and conversational expectations (Shaikh et al., 2023; Zheng et al., 2024; Carta et al., 2023; Zhang et al., 2024; Peng et al., 2023; Li et al., 2022, 2024). In general, the closeness of LM response to those that might be generated naturally/by a human is typically used to reason about the "grounding" of a language model. Work in this domain can include evaluating how accurate responses are based on provided contexts (Lee et al., 2023), identifying gaps in generation vs human responses (Shaikh et al., 2023) and improving models to better perform in interactive environments (Carta et al., 2023)

3.5 Knowledge Grounding

Closely linked to language model grounding, knowledge grounding refers to integrating structured or unstructured external knowledge into machine learning models to improve factual consistency and provide outputs based on real-world facts and reliable data. This is commonly applicable in Question Answering (QA) systems which could be vision/video (Xiao et al., 2024; Chen et al., 2022, 2023a; Urooj et al., 2021; Shrestha et al., 2020; Khan et al., 2022) or language based (Shi et al., 2024), retrieval and retrieval-augmented systems (Hannan et al., 2023; Shi et al., 2024), reasoning tasks (Zhang et al., 2021; Zhu et al., 2024; Kottur et al., 2019) and conversational AI (Ahuja et al., 2022; Liu and Chai, 2015; Wu et al., 2021; Gao et al., 2022). Most of the work in this series involves the creation and validation of datasets as large amounts of data are required for effective knowledge grounding, and may include video, audio, text, or even programmatic (Wang et al., 2023) data cues to ground models.

3.5.1 Common Grounding

Common grounding, the "process of creating, repairing and updating mutual understandings" (Udagawa and Aizawa, 2020, 2019) is a subset of knowledge grounding in conversational contexts. It is aimed at aligning different sources of information (such as language, perception, or actions) so that entities (whether human or machine) can interact with each other meaningfully and consistently.

3.6 Grounding in Reinforcement Learning (RL)

Grounding in RL ensures that an agent's decision-making is aligned with real-world feedback by refining policies through interaction (Carta et al., 2023) instead of data based solutions, as summarized in the previous section. This could include functional grounding for agents to iteratively update their policies, language grounding in multi-agent settings to help agents decompose tasks into sub-goals and coordinate actions based on natural language instructions (Ding et al., 2023), or temporal grounding to localize events sequentially (He et al., 2019). Grounding enhances RL's effectiveness in robotics, embodied AI, and language-conditioned agents.

4 Discussion

We now discuss the insights uncovered in the summarization in Sec3. We note that although we have uncovered five distinct senses, with multiple sub-senses, there is a high degree of correlation among them as well. Below, we highlight and sum up the similarities and differences in the different senses.

4.1 Similarities Across Different Senses of Grounding

4.1.1 Linking Information Across Modalities

All forms of grounding involve associating data across different sources, whether it's text and vision (visual grounding), language and knowledge (knowledge grounding), or perception and action (embodied visual grounding).

4.1.2 Enhancing AI Interpretability and Accuracy

All the grounding methods are aimed at improving machine learning models, specifically to ensure that generated outputs align with real-world contexts or user expectations.

4.1.3 Data Dependency

Most grounding techniques, especially knowledge grounding and visual grounding, require large datasets to train models effectively. In fact, a significant amount of papers analyzed were aimed at creating or evaluating various datasets and accompanying benchmarks.

4.2 Key Differences Across Different Senses of Grounding

4.2.1 Type of Grounding Objective

This is where the senses primarily differentiate. We can categorize the objectives largely into three categories:

1. Semantic Grounding (Meaning-Based).

   This is most relevant to language model grounding and knowledge grounding, where the goal is to ensure factual consistency and semantic coherence.

2. Perceptual Grounding (Perception-Based).

   This is most relevant to visual grounding, 3D grounding, and multi-modal grounding, where the system must associate sensory data (e.g., images, video, sounds) with linguistic descriptions.

3. Actionable Grounding (Action-Based).

   This is most relevant to embodied grounding, which focuses on grounding perception to actions in real-world tasks, such as robotics or interactive AI agents.

4.2.2 Real-Time vs. Static Grounding

Another difference is whether the grounding needs to apply to real-time decision making or to enhance static learning. Visual grounding, especially embodied or video grounding, usually involves real-time decision making and continuous updates based on interactions with the environment, while knowledge and language model grounding may involve providing the data to the AI system statically, from which it then learns to ensure factual soundness.

4.2.3 The Role of External Knowledge

Some forms of grounding, e.g. knowledge or language model grounding, make use of external knowledge that is integrated into the model to ground themselves. On the contrary, methods such as visual or spatial-temporal grounding, which observe and learn from data, likely won't use external knowledge.

5 Extended Analysis of the Use and Evolution of "Grounding"

In this section, we answer the questions outlined in Sec2.4 using quantitative analysis. We keep this analysis separate from the senses identified and look at the broader effect of the term "grounding" itself.

5.1 How has the presence of "grounding" changed in papers over-time?

We plot the count by year of all papers that contained any occurrence of the word "grounding" in the AI-paper-crawl corpus (Fig 1). There is a sharp increase in the presence of "grounding" in relevant literature in the late 2010s - early 2020s. Although this might be in part due to the increase in the number of publications in general, a similar trend is also seen in our smaller corpus of selected papers (Fig 2). This suggests that the term has been gaining traction in recent years, with more and more people acknowledging its importance in the AI/ML context.

5.2 How does the presence of related keywords change in papers over time? Which words are more common?

We compile a list of 24 keywords commonly associated with grounding (Appendix C). For this list, we find trends in frequency (how often each word occurs in papers (Fig 4) as well as the number of pa-

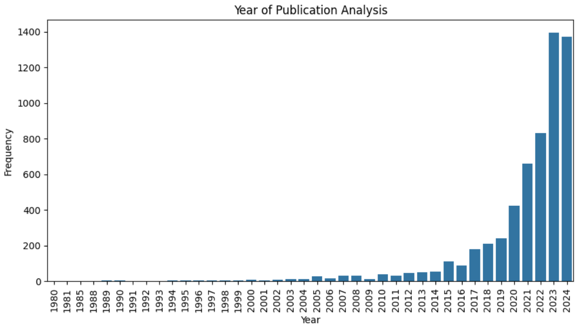

Figure 1: Yearly count of all papers with the word "grounding" in them

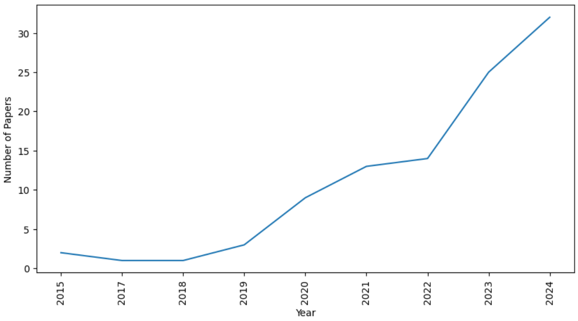

Figure 2: Yearly trend of paper counts for selected papers

pers it comes up in (Fig 3) and yearly occurrences (Fig 5).

We find that most words are infrequent in our corpus. The words that are frequent are unsurprising - they largely align with the senses identified in the previous section (visual, video, temporal, multimodal). The high frequency of "visual" particularly checks out with the largest number of papers and the highest number of sub-senses analyzed for "visual grounding" in the analysis in the previous sections. However, the keywords "contrastive" and "symbol", also came up as frequent in this analysis, which is interesting as they were not uncovered in our literature review and summarization of the different senses. Figure 4 explains this to some extent: while these keywords are present in a large number of papers, they are not mentioned frequently enough as compared to some of the other keywords that made up our sense categorization. This begets the need for more fine-grained analysis on how these terms are used in the papers and why they occur repeatedly, something that can be targeted in future work.

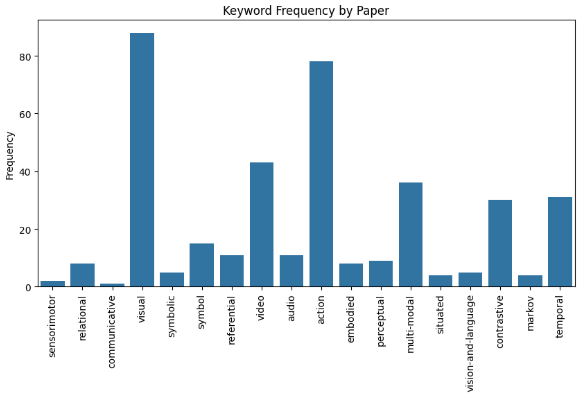

Figure 3: Number of papers containing the keyword for selected papers

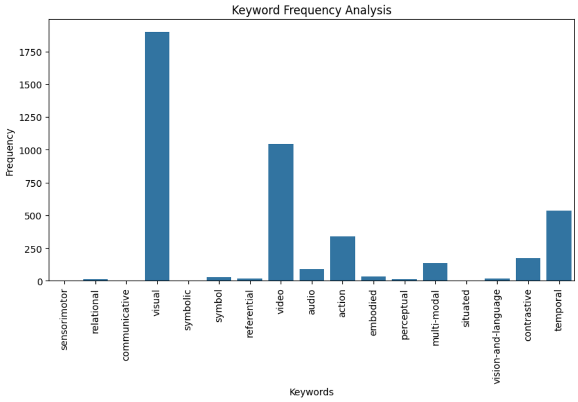

Figure 4: Frequency of keywords over all papers for selected papers

5.3 Which ML sub-fields are commonly mentioned in these papers, are there any temporal trends?

Finally, we compile a list of sub-fields commonly associated with grounding (Appendix D) and evaluate yearly trends for the occurrences of these terms in the selected papers.

Our analysis showed that not all the sub-fields we added were significant, and several showed up in less than 5 papers. Discarding these sub-fields, we plotted a temporal trend of the use of each sub-field in grounding literature over the years (Fig 6).

The plot uncovers distinct trends in the occurrence of sub-fields. Computer vision and reinforcement learning both see a dip as multimodal and object detection peak in recent years, which seem to align with the direction academia and industry have been taking recently.

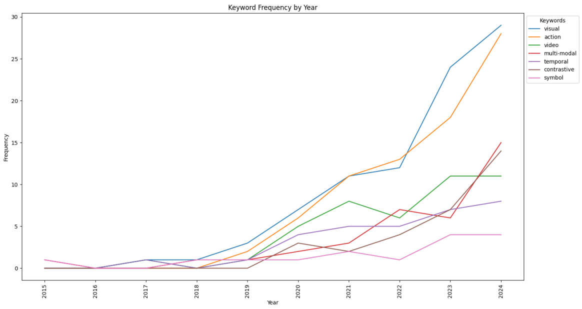

Figure 5: Yearly trend of keywords for selected papers

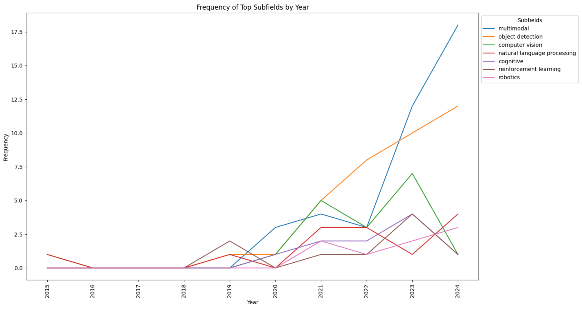

Figure 6: Yearly trend of sub-field occurrence in selected papers

6 Limitations

Our work has several limitations due to its restricted scale. A limitation of our methodology is that we do not consider other forms of the word "grounding" when shortlisting papers. There may be papers that only use variations like "grounded" or "ground", which means that we might have possibly missed considering certain significant senses. Similarly, making arbitrary choices for the number of papers selected and the number of topics and clusters for data analysis is also something that can be improved in future works - these should be based on conventions in the ML/AI community or by the size of the corpus. Our choice of words for the set of keywords and sub-fields used to extract different insights from the corpus was arbitrarily made based on the authors' domain knowledge as well and could be more methodologically chosen, for example, from a secondary data source. We also do not conduct our analysis by conference to limit the scale of our work, which can introduce biases and also result in missing important trends. Extensions of this work can target these limitations.

7 Conclusion

Despite methodological limitations, our work provides a foundational framework for understanding "grounding" in AI/ML. By delineating its diverse interpretations and contextual applications, this study contributes to clearer discourse and clarification of the term within the research community.

Future work can target the identified limitations and extend the analysis to examine the historical evolution of "grounding" concepts over time, use current trends to predict future directions, and assess how various grounding techniques influence the performance and reliability of AI/ML models across diverse applications, offering actionable guidance for researchers and practitioners aiming to implement grounding effectively.

References

Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. 2023. Ht-step: Aligning instructional articles with how-to videos. *Advances in Neural Information Processing Systems*, 36:50310–50326.

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. 2022. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20566–20576.

Peijun Bao, Qian Zheng, and Yadong Mu. 2021. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 920–928.

Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbron, and Bernard Ghanem. 2023. Localizing moments in long video via multimodal guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13667–13678.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16464–16473.

Cambridge. 2025. Grounding - cambridge dictionary. [Online; accessed 2025-03-28].

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, pages 3676–3713. PMLR.

Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. 2024. What when and where? self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18419–18429.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023a. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325.

Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. 2023b. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18109–18119.

Ziluo Ding, Wanpeng Zhang, Junpeng Yue, Xiangjun Wang, Tiejun Huang, and Zongqing Lu. 2023. Entity divider with language grounding in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 8103–8119. PMLR.

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. Unigdd: A unified generative framework for goal-oriented document-grounded dialogue. *arXiv preprint arXiv:2204.07770*.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. 2023. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383.

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. 2023. Rgnet: A unified clip retrieval and grounding network for long videos. *arXiv preprint arXiv:2312.06729*.

Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8393–8400.

Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523.

Wenhui Jiang, Yibo Cheng, Linxin Liu, Yuming Fang, Yuxin Peng, and Yang Liu. 2024. Comprehensive visual grounding for video description. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2552–2560.

Yang Jin and Yadong Mu. 2024. Weakly-supervised spatio-temporal video grounding with variational cross-modal alignment. In *European Conference on Computer Vision*, pages 412–429. Springer.

Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Vitoria Lobo, and Mubarak Shah. 2022. Weakly supervised grounding for vqa in vision-language transformers. In *European Conference on Computer Vision*, pages 652–670. Springer.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Hyunji Lee, Sejune Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, and Minjoon Seo. 2023. How well do large language models truly ground? *arXiv preprint arXiv:2311.09069*.

Phillip Y Lee and Minhyuk Sung. 2024. Reground: Improving textual and spatial grounding at no cost. In *European Conference on Computer Vision*, pages 275–292. Springer.

Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, and 1 others. 2022. End-to-end modeling via information tree for one-shot natural language spatial video grounding. *arXiv preprint arXiv:2203.08013*.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, and 1 others. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Xiangru Lin, Guanbin Li, and Yizhou Yu. 2021. Scene-intuitive agent for remote embodied visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7036–7045.

Changsong Liu and Joyce Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. 2021. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6032–6041.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded visual captions without localization supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 353–370. Springer.

Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, and Matt Gardner. 2022. A meta-framework for spatiotemporal quantity extraction from text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2736–2749.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Seed42Labs. Seed42lab/ai-paper-crawl · datasets at hugging face. [Online; accessed 2025-03-29].

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding gaps in language model generations. *arXiv preprint arXiv:2311.09144*.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Zhengliang Shi, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. *arXiv preprint arXiv:2406.14891*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2020. A negative case analysis of visual grounding methods for vqa. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8172–8181.

Sijie Song, Xudong Lin, Jiaying Liu, Zongming Guo, and Shih-Fu Chang. 2021. Co-grounding networks with semantic attention for referring expression comprehension in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1355.

Penglei Sun, Yaoxian Song, Xinglin Pan, Peijie Dong, Xiaofei Yang, Qiang Wang, Zhixu Li, Tiefeng Li, and Xiaowen Chu. 2024. Multi-task domain adaptation for language grounding with 3d objects. In *European Conference on Computer Vision*, pages 387–404. Springer.

Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Takuma Udagawa and Akiko Aizawa. 2020. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9081–9089.

Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. 2024. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *European Conference on Computer Vision*, pages 196–213. Springer.

Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. 2021. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8465–8474.

Ning Wang, Jiajun Deng, and Mingbo Jia. 2024. Cycle-consistency learning for captioning and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5535–5543.

Renhao Wang, Jiayuan Mao, Joy Hsu, Hang Zhao, Jiajun Wu, and Yang Gao. 2023. Programmatically grounded, compositionally generalizable robotic manipulation. *arXiv preprint arXiv:2304.13826*.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2024. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and 1 others. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14085–14093.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. 2024. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238.

Yiming Zhang, ZeMing Gong, and Angel X Chang. 2023. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236.

Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. 2021. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. *Advances in Neural Information Processing Systems*, 34:18513–18526.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. 2021. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8454.

Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. 2023. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23151–23160.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer.

Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. 2024. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pages 151–168. Springer.

A Use of Generative AI tools

We acknowledge that we have used the generative AI tool ChatGPT to revise some sections of the paper and for initial brainstorming. We also used Microsoft Copilot to streamline some of the scripts used in the data analysis.

Conference	**Num Papers**
AAAI	11
ACL	9
CVPR	31
ECCV	18
EMNLP	2
ICCV	10
ICLR	3
ICML	3
IJCAI	5
NAACL	4
NIPS	4

Table 1: Number of papers in the corpus from each conference

B Details of the Selected 96 Papers

Table 1 shows the distribution of conferences in the selected paper corpus. Table 2 shows the distribution of common keywords in the corpus. The 7 most common keywords are:

- visual
- action
- video
- multi-modal
- temporal
- contrastive
- symbol

C Keywords Associated with Grounding

The following set of words were arbitrarily chosen based on background knowledge and brief review of relevant literature. A more methodological way of compiling this list could be considered in future work.

> "sensorimotor", "relational", "communicative", "visual", "symbolic", "symbol", "referential", "video", "audio", "action", "embodied", "perceptual", "multi-modal", "situated", "vision-and-language", "contrastive", "markov", "temporal"

D Common AI/ML Sub-fields

The following set of fields were arbitrarily chosen based on background knowledge and brief review of relevant literature. A more methodological way of compiling this list could be considered in future work.

Keyword	Num Papers
sensorimotor	2
relational	8
communicative	1
visual	88
symbolic	5
symbol	15
referential	11
video	43
audio	11
action	78
embodied	8
perceptual	9
multi-modal	36
situated	4
vision-and-language	5
contrastive	30
markov	4
temporal	31

Table 2: Keyword frequency in the corpus

"robotics", "natural language processing", "computer vision", "reinforcement learning", "explainable AI", "XAI", "multimodal", "cognitive", "language interpretation", "language generation", "dialogue systems","human-robot interaction", "deep learning", "healthcare", "medical imaging", "autonomous vehicles", "affective computing", "emotion recognition", "speech reconition", "speech synthesis", "audio generation", "video generation", "image generation", "image recognition", "image classification", "image segmentation", "object detection", "object recognition", "object tracking"