Limited Linguistic Diversity in Embodied AI Datasets

Selma Wanna[†] [1], Agnes Luhtaru[†] [2], Jonathan Salfity[3], Ryan Barron[1], Juston Moore[1], Cynthia Matuszek[4], Mitch Pryor[3],

[1]Los Alamos National Laboratory, Los Alamos, USA,
[2]Institute of Computer Science, University of Tartu, Estonia,
[3]Department of Mechanical Engineering, The University of Texas at Austin, USA,
[4]University of Maryland, Baltimore County, Maryland, USA

Equal contribution: [†] denotes equal contribution; **Correspondence:** slwanna@lanl.gov, agnes.luhtaru@ut.ee

Abstract

Language plays a critical role in Vision-Language-Action (VLA) models, yet the linguistic characteristics of the datasets used to train and evaluate these systems remain poorly documented. In this work, we present a systematic dataset audit of several widely used VLA corpora, aiming to characterize what kinds of instructions these datasets actually contain and how much linguistic variety they provide. We quantify instruction language along complementary dimensions—including lexical variety, duplication and overlap, semantic similarity, and syntactic complexity. Our analysis shows that many datasets rely on highly repetitive, template-like commands with limited structural variation, yielding a narrow distribution of instruction forms. We position these findings as descriptive documentation of the language signal available in current VLA training and evaluation data, intended to support more detailed dataset reporting, more principled dataset selection, and targeted curation or augmentation strategies that broaden language coverage.

1 Introduction

With advances in large language models (LLMs) and multimodal learning, language is increasingly used as an input across research fields, enabling practical systems. In robotics, this is reflected in the growing focus on Vision-Language-Action (VLA) models such as OpenVLA (Kim et al., 2025), RT-X (Collaboration et al., 2024), and π0.5 (Intelligence et al., 2025). Much of this progress has been driven by datasets like Open X-Embodiment (OXE, Collaboration et al., 2024), which are larger and more diverse across objects, scenes, and embodiments than earlier robotics datasets, supporting a shift toward end-to-end generalist robotic systems that use language to specify tasks.

Despite this progress, language—while a core modality in VLA systems—is often overlooked in dataset documentation and evaluations. Datasets in the OXE collection were not originally created for generalist VLA training, and documentation emphasizes diversity across objects, scenes, and embodiments, with little discussion of instruction language (Collaboration et al., 2024). Meanwhile, a growing body of work reports limited semantic robustness of VLA models despite reliance on LLM backbones, including sensitivity to paraphrases, performance drops in the presence of distractor objects, and related failures (Gao et al., 2025; AgiBot-World-Contributors et al., 2025; Wang et al., 2024). These generalization issues suggest that language understanding may be underemphasized in current VLA development; documenting the linguistic features of the data that shape model performance is a crucial step for future research.

To address this, we analyze instruction language in several VLA datasets from OXE and compare it to other robotics datasets and common instruction-tuning corpora used for LLM training. We characterize language coverage across three dimensions: lexical redundancy, semantic diversity, and structural diversity (cf. Section 3). We find that current VLA datasets exhibit a narrow and repetitive

RT-1

move sponge near orange can
move blue plastic bottle near sponge

BridgeData

pick the carrot and place it inside the stainless steel pot
pick the strawberry and put it into the stainless pot.

Language Table

slide the red star into the cube
slightly move the red circle below the red star

Figure 1: Examples of language instructions from selected VLA datasets in the OXE collection, illustrating limited linguistic diversity. Colored tokens denote verbs (purple), object nouns (pink), spatial relations (teal), and descriptive adjectives (yellow)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11058–11086

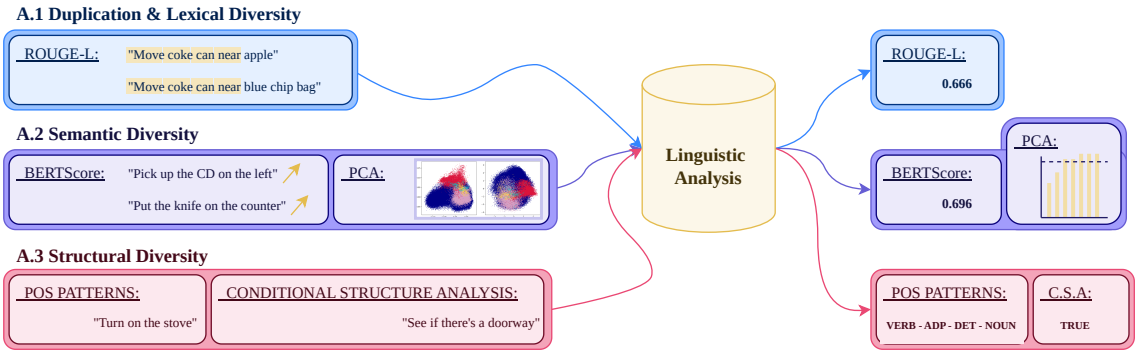


Figure 2: We analyze linguistic diversity in embodied AI datasets across three categories: Analysis 1 (A.1) Duplication & Lexical diversity, Analysis 2 (A.2) Semantic diversity, and Analysis 3 (A.3) Structural diversity.

instruction distribution: fewer than 2% of instructions are unique, lexical diversity is limited relative to the comparison datasets, and structural forms are often templated (cf. Section 5). Beyond repetition, linguistically richer constructions are largely missing: across all studied datasets, fewer than 1% of commands contain negation, and conditionals are similarly rare.

Overall, we provide a systematic characterization of language used in VLA datasets (see Figure 1 for command examples). By making these properties explicit, this language-focused documentation can inform synthetic data generation and future data collection, help contextualize reported generalization gaps such as sensitivity to paraphrases, and guide the design of new evaluation protocols.

2 Background

Many recent VLA models build on modern vision–language models (VLMs) that combine a vision encoder and an LLM via adapters (e.g., PaliGemma) (Beyer et al., 2024). Action generation is typically added either autoregressively, by extending the LLM vocabulary with action tokens (Kim et al., 2024; Pertsch et al., 2025), or via a separate action expert trained with diffusion/flow-matching objectives (Black et al., 2024).

A key challenge is integrating action supervision without catastrophic forgetting or degrading pretrained VLM capabilities, which lead to semantic generalization failures. Prior work studies how action representations affect language generalization (Grover et al., 2025) and how combining an autoregressive VLM with an action expert can change or degrade VLM behavior (Driess et al., 2025). Other works investigate co-training with multimodal understanding data (Zhou et al., 2025; Gao et al., 2025; Grover et al., 2025), or incorporate reasoning-style supervision, resulting in slower inference (Zawalski et al., 2024; Chen et al., 2025).

In contrast, there is comparatively limited work on what training-data properties support effective VLA grounding. Prior results in vision–language grounding suggest that coverage and data quality are critical for alignment, and that gaps in coverage are associated with failures even on simple capabilities, with performance degrading on underrepresented concepts and improving with targeted coverage (Udandarao et al., 2024; Zhang et al., 2024). For VLA models, however, dataset analyses have mostly focused on non-linguistic sources of shortcuts (e.g., fragmentation, viewpoints, backgrounds); Xing et al. (2025) highlights such issues and briefly notes limited language variation, but does not provide a detailed characterization of instruction language.

While these findings do not isolate linguistic diversity as a causal factor in VLA performance, they suggest that gaps in training data coverage can limit generalization and language–action grounding even when the underlying backbone has strong capabilities, motivating the need to study linguistic variability in VLA datasets.

3 Measuring Linguistic Diversity for EAI

There is no widely adopted standard for characterizing linguistic diversity in embodied AI (EAI) datasets. Dataset *diversity* is inherently vague, and many studies claim to address diversity without clearly defining what it entails (Zhao et al., 2024). Linguistic diversity itself spans multiple dimensions: for example, Guo et al. (2024) discusses lexical diversity (variation in word choice), semantic diversity (variation in meaning), and syntactic di-

versity (variation in grammatical structure). Tevet and Berant (2021) distinguishes between form diversity, which relates to surface-level variation, and content diversity, which captures deeper semantic differences.

In VLA datasets, however, diversity is typically defined in non-linguistic terms, such as the number of distinct objects, environments, or tasks (Collaboration et al., 2024). While this framing may reflect aspects of semantic variation (e.g., which actions are performed on which objects), it provides little insight into the diversity of natural-language instructions, including variation in word choice, grammatical structure, and meaning expression.

We construct a diversity evaluation that incorporates a broad range of analysis metrics, largely following the categorization followed by Tevet and Berant (2021). Our evaluation covers three analyses (see Figure 2):

- **Analysis 1: Duplication and Lexical Diversity**, including standard statistics (e.g., number of unique unigrams, sentence length, lexical overlap) and common diversity metrics;
- **Analysis 2: Semantic Diversity**, measured using embedding based methods, and verb–direct object–adverbial diversity;
- **Analysis 3: Structural Diversity**, including syntax (POS-patterns) and higher-level linguistic phenomena, like negations.

Each analysis targets a distinct aspect of linguistic diversity, enabling a more comprehensive understanding of the textual variation present in datasets used to train models. In this work, we focus specifically on the repetitiveness and uniformity of language in VLA datasets, rather than on diversity related to sociocultural or demographic bias. Below, we describe each analysis in detail. The corresponding results are presented in Section 5. We discuss additional qualitative patterns in selected datasets in the Appendix A.

3.1 Duplication and Lexical Diversity

Many OXE datasets were created primarily for imitation learning, where language is often treated as an auxiliary signal. Consequently, datasets built from human-teleoperated or scripted demonstrations frequently exhibit substantial repetition in instruction phrasing and limited coverage of distinct command types. In the LLM literature, duplication is associated with increased memorization and reduced reliance on generalization (Kandpal et al., 2022; Lee et al., 2022). More broadly, deep networks have sufficient capacity to fit random labels, highlighting that memorization can be easy in overparameterized settings (Zhang et al., 2017). We hypothesize that similar risks may arise for VLA models trained on highly repetitive instructions.

In discussions of linguistic diversity—especially in the context of LLM-generated text—lexical diversity is typically the primary focus. Metrics such as BLEU (Papineni et al., 2002; Zhu et al., 2018) and ROUGE (Lin, 2004), although originally designed for evaluation of generation quality, have been adapted to approximate diversity. Similarly, compression ratio (CR), based on gzip, has been used as a proxy for textual diversity and has been shown to be effective at distinguishing LLM-generated from human-authored text (Shaib et al., 2025).

To assess duplication and lexical diversity, we compute basic linguistic statistics—such as the number of unique commands, sentences, and unigrams—as well as five textual diversity metrics. ROUGE-L and compression ratio (CR), following Shaib et al. (2025), are reported in Table 2 while the remaining three metrics: BLEU, Jaccard similarity and Levenshtein distance are provided in Table 4.

3.2 Semantic Diversity

Semantic diversity is typically measured using embedding-based metrics that group paraphrases closely in semantic space. While sentence embeddings such as BERTScore (Zhang* et al., 2020) often struggle to capture aspects like syntax, antonymy, and word order (Zhang et al., 2023; Mahajan et al., 2024), they can be used for evaluating variation in content (i.e., what is said, not how) (Tevet and Berant, 2021; Stasaski and Hearst, 2022). In the context of robotics, such metrics can help quantify the variety of actions and object interactions described in natural language instructions. Among the dimensions we evaluate, semantic diversity is expected to correlate most closely with the number of distinct objects and skills in the dataset, as these reflect the variety of task meanings that natural language instructions encode.

Motivated by this, we include BERTScore as a pairwise diversity metric between individual instructions. In addition, we apply Principal Component Analysis (PCA) to sentence embeddings of entire datasets and report the number of components

required to explain 95% of the cumulative variance (Fan et al., 2010; Verleysen and Lee, 2013), to capture dataset-level variation in instruction semantics. For sentence embeddings, we use four common encoders - USE (512D) (Cer et al., 2018), SBERT (768D) (Reimers et al., 2019), CLIP (512D) (Radford et al., 2021), and SONAR (1024D) (Duquenne et al., 2023). For brevity, we report USE in Table 2 but provide the results across all encoders in Table 5. We further justify this approach in the Appendix D.

In addition to these metrics, we examine verb, direct object, and adverb diversity to provide more interpretable and domain-specific insights for robotics. Specifically, we assess how many distinct verbs are used with each direct object in manipulation datasets. In contrast, direct object structures are less relevant for navigation-focused datasets. For these, the manner in which an instruction is followed—such as the use of directional terms (e.g., "north," "forward"), location-based modifiers (e.g., "around," "inside"), and manner descriptors (e.g., "slowly," "directly")—is more pertinent. Low counts of these features suggest limited interaction diversity, which could introduce biases. If a model has learned to perform only specific actions with certain objects, it may eventually learn to ignore the language command, as neural networks have a well-known simplicity bias (Shah et al., 2020).

3.3 Structural Diversity

We group both syntactic variation and the presence of higher-level linguistic phenomena—such as negation, conditionals, cycles, and multi-step instructions—under the category of structural diversity. Prior work has shown that limited syntactic variety can amplify model biases, whereas incorporating more diverse grammatical structures improves generalization and reduces overfitting to shallow heuristics (Aggarwal et al., 2022). Common measures of syntactic diversity include part-of-speech (POS) patterns and constituency or dependency parse trees. In our evaluation, we include the frequency of different POS patterns and use constituency tree similarity (Moschitti, 2006) to quantify syntactic variation across datasets (reported in Table 2).

Beyond surface-level syntax, structural diversity also includes compositional and logic-oriented constructions. Logical structures can support reasoning capabilities in language models (Uchiyama et al., 2024), and challenges in handling negation remain prominent in NLU tasks (Hossain et al., 2022) and Zhang et al. (2025) demonstrate that VLMs struggle significantly with negation. In the context of robotics, the presence of linguistic phenomena such as negation, conditionals, and multi-step instructions is particularly important. Even if these structures do not strongly affect current VLA benchmark scores, they remain important because they express constraints, exceptions, contingency, and sequential logic that arise naturally in real-world interaction. These structures increase the complexity of commands that a robot must understand and execute. Without them, instructions are limited to simple, atomic actions, reducing the system's ability to interpret nuanced or context-dependent behaviors. For example, instructions such as *"give me the apple that is not rotten"*, *"if you have picked up the apple, wash it"*, or *"pick up the apple and place it on the cutting board"* require compositional understanding and the ability to process conditional logic, negation, and sequential actions. Capturing such structures in training data may also support reasoning and help models handle more complex commands.

To quantify this aspect of structural diversity, we estimate the proportion of instructions that contain negation, conditionals, cycles, or multi-step commands. We rely on syntactic cues from dependency parses, specific keyword patterns, and part-of-speech tags to identify these phenomena. In addition, we manually annotate a subset of each dataset to validate the accuracy of the automatic heuristics and to better characterize the linguistic structures present.

4 Datasets

To investigate the linguistic characteristics of data used to train VLA models, we analyze a set of datasets drawn primarily from the OXE collection, which are among the most widely used sources for VLA training (Kim et al., 2024; Collaboration et al., 2024; Intelligence et al., 2025). OXE contains over 40 datasets with language annotations; we therefore focus on a representative subset chosen to balance (i) relevance/usage in the literature (e.g., citation frequency and common inclusion in training or selection pipelines), (ii) scale (sufficient episodes for stable estimates), and (iii) coverage of distinct linguistic regimes (templated vs. natural, open-ended instructions). An overview is provided in Table 1, with further details in Section 4.1.

Dataset	Citations	Focus	Language Style
Instruction-Tuning			
OASST2 (Köpf et al., 2023)	779+	LLM instruction tuning	Crowdsourced instructions
Alpaca (Taori et al., 2023)	4361+	LLM instruction tuning	Generated from seed data
LLaVA-Instruct (Liu et al., 2023b)	7286+	VLM instruction tuning	Generated, specific questions
Language-Focused Robotics Datasets			
ALFRED (Shridhar et al., 2020)	936+	Household task instruction following	Step-by-step, high-level
SCOUT (Lukin et al., 2024)	6+	Two-way, task-oriented dialogue	Unconstrained, interactive
VLA Datasets			
Open X-Embodiment (Collaboration et al., 2024)	746+	Collection of datasets	Varied, not always included
RT-1 (Brohan et al., 2023)	1693+	Kitchen instruction following	Concise, imperative, templated
BRIDGE (Ebert et al., 2022)	330+	Skill generalization across domains	Diverse, step-by-step
TacoPlay (Rosete-Beas et al., 2022)	108+	Task-agnostic "play" behaviors	Simple, low-variety, templated
Language Table (Lynch et al., 2023)	290+	Open-vocab spatial manipulation	Natural, open-ended
LIBERO (Liu et al., 2023a)	450+	Knowledge transfer in robot learning	Natural[1]

Table 1: Overview of the datasets explored in this work. We include citation counts for each dataset; note that some of the referenced works focus primarily on dataset creation, while others introduce new methods alongside the dataset. Citation data from Semantic Scholar (Alpaca from Google Scholar).

As many of the computed metrics are difficult to interpret in isolation, we include additional reference datasets to contextualize the results: robotics datasets that are not typically used to train generalist VLA models but place greater emphasis on language, and (in selected experiments) instruction-tuning datasets from outside robotics. These references are not intended to define "ideal" properties; rather, they serve as anchors for interpreting the linguistic characteristics of existing VLA training corpora. In non-robotics instruction-following datasets, "commands" refer to individual sentence examples.

4.1 VLA Datasets

From the OXE collection, we analyze four widely studied datasets—RT-1 (Brohan et al., 2023), BRIDGE (Ebert et al., 2022), TacoPlay (Rosete-Beas et al., 2022), and Language Table (Lynch et al., 2023). We include RT-1 (Brohan et al., 2023) because it is frequently cited, was introduced alongside the model, is among the largest OXE datasets in terms of episodes, and is also often selected by automatic data selection methods (Hejna et al., 2025; Dass et al., 2025); its instructions are primarily imperative and highly templated. To contrast this with more natural language, we include Language Table (Lynch et al., 2023), which has the most episodes in OXE and targets open-vocabulary spatial manipulation in controlled tabletop environments with more open-ended, dialogue-like interactions. We further include BRIDGE (Ebert et al., 2022) as a complementary generalization-oriented dataset: while it also aims to support broad task generalization, it covers a wider range of tasks and environments, supports tool use and fine-grained object interactions, and is commonly featured in generalization evaluations.[2] Finally, we include TacoPlay (Rosete-Beas et al., 2022) to represent a task-agnostic "play" paradigm learned from unstructured interaction data; although its language remains largely templated.

We additionally include LIBERO (Liu et al., 2023a), a simulation-based benchmark focused on knowledge transfer. While not part of the OXE collection, LIBERO has been increasingly adopted in recent experimental and evaluation setups.

4.2 Language-Focused Robotics Datasets

We also include two robotics datasets that prioritize language interaction, although they are not primarily used to train generalist VLA models. ALFRED (Shridhar et al., 2020) emphasizes natural language through fine-grained, step-by-step instructions aligned with low-level actions, making it well-suited for studying task decomposition and instruction-following.

SCOUT (Lukin et al., 2024) contains the most spontaneous, dialogue-based interaction data among the datasets considered. It captures unconstrained human-robot communication in navigation scenarios, supporting adaptive, context-aware exchanges beyond static command formats. SCOUT includes transcriptions from real-world robot operators and provides detailed linguistic statistics. For our analysis, we focus on utterances from the robot commander role to align with the style of other datasets.

[2]We use the original Bridge dataset from the OXE download link. A newer version, which is now more commonly used, contains more episodes (Walke et al., 2023).

Dataset	A.1 Duplication & Lexical					A.2 Semantic		A.3 Structural
	# Sent	# Uniq (% Uniq)	# Words	CR ↓	ROUGE-L ↓	BERTScore ↓	USE ↑	Tree Kernel ↓
Instruction-Tuning Datasets								
OASST2 (Köpf et al., 2023)	42K+	39,301 (93.33%)	**35,445**	**2.75**	**0.05 ± 0.00**	**0.45 ± 0.00**	**254**	**2.25 ± 0.01 %**
Alpaca (Taori et al., 2023)	53K+	52,996 (**99.81%**)	18,141	3.20	0.10 ± 0.00	0.57 ± 0.00	231	3.66 ± 0.05 %
LLaVA-Instruct (Liu et al., 2023b)	**366K+**	**261,892** (71.45%)	15,477	4.41	0.21 ± 0.00	0.61 ± 0.00	184	7.46 ± 0.13 %
Language-Focused Robotics Datasets								
ALFRED (Shridhar et al., 2020)	**162K+**	**126,005** (**79.9%**)	**2,627**	5.91	0.21 ± 0.00	0.64 ± 0.00	**159**	5.71 ± 0.14 %
SCOUT (Lukin et al., 2024)	23K+	8,795 (39.4%)	1,631	**4.85**	**0.07 ± 0.00**	**0.49 ± 0.00**	148	**1.89 ± 0.22 %**
VLA Datasets								
RT-1 (Brohan et al., 2023)	3.7M+	577 (0.02%)	49	118.20	0.19 ± 0.01	0.64 ± 0.00	33	5.09 ± 0.20 %
BRIDGE (Ebert et al., 2022)	864K+	11,693 (1.4%)	**1,189**	64.90	**0.15 ± 0.00**	**0.60 ± 0.00**	**125**	**3.68 ± 0.12 %**
TacoPlay (Rosete-Beas et al., 2022)	214K	403 (0.2%)	74	158.86	0.30 ± 0.01	0.68 ± 0.00	42	8.86 ± 0.13 %
LanguageTable (Lynch et al., 2023)	**7.0M+**	**127,370** (**1.81%**)	928	**56.64**	0.29 ± 0.00	0.70 ± 0.00	86	9.19 ± 0.14 %
LIBERO (Liu et al., 2023a)	6.5K	112 (1.72%)	79	134.86	0.38 ± 0.00	0.71 ± 0.00	34	12.22 ± 0.29 %

Table 2: Summary of all sentences (# Sent), unique sentences (% Uniq, # Uniq), unigrams (# Words), and text similarity measures across various datasets. Pairwise scores (ROUGE-L, BERTScore, Tree Kernel) are computed by sampling 1,000 commands from each dataset, repeated three times for robustness. Arrows indicate increasing linguistic diversity. CR stands for Compression Ratio. The Tree Kernel method is from Moschitti (2006). USE refers to the minimum # of PCA components derived from USE embeddings to explain 95% variance for each dataset. The arrow points towards increased diversity.

4.3 Instruction-Following Datasets

To contextualize the language complexity of modern robotics datasets, we include two widely used text-only instruction-tuning datasets. To incorporate human-authored instructions, we use the English portion of the Open Assistant Conversations Dataset (OASST2) (Köpf et al., 2023). As a comparison set, we also include Alpaca (Taori et al., 2023), which contains LLM-generated instructions. To represent language from visual instruction tuning, we add LLaVA-Instruct (Liu et al., 2023b), a dataset used to train vision-language models commonly employed in VLA systems. LLaVA-Instruct primarily consists of questions about image content—such as object types, counts, actions, locations, and spatial relationships—resulting in more constrained language patterns. For all instruction-tuning datasets, we extract only the instruction texts, discarding the associated inputs and model responses.

5 Results

In this section, we highlight key results of command duplication and linguistic diversity—lexical, semantic, and structural—across a range of datasets. Quantitative results allow us to identify recurring characteristics that shape the language used in embodied AI benchmarks. We focus on contrasts between VLA datasets, other robotics datasets, and instruction-tuning corpora, exploring how linguistic patterns vary across these domains. All quantitative results, specific implementation details (e.g., text preprocessing, POS-tag extraction), and additional examples are provided in the Appendices B-G.

Analysis 1: Duplication. Language command duplication is common across all the OXE datasets we analyzed, as well as in LIBERO. Our analysis (see Table 2) reveals a notable disparity: in VLA datasets, fewer than 2% of language instructions contain unique wording, which contrasts sharply with other robotics datasets and instruction-following corpora. This is largely due to the same command being paired with multiple action trajectories via multiple trials. Outside of VLA datasets, the lowest percentage of unique sentences appears in SCOUT, with around 40% uniqueness. Many SCOUT utterances are very short—on average fewer than five words (see Figure 6 in the Appendix C.1)—which likely contributes to higher repetition compared to other non-VLA datasets. The low percentage of unique commands in OXE datasets suggests that the number of recorded episodes is not indicative of the diversity of natural language commands and highlights that during training the models frequently see the exact same commands.

Analyses 1–3: Lexical, Semantic, and Structural Diversity Metrics. Overall, the textual diversity metrics align: VLA datasets exhibit the lowest diversity across lexical, semantic, and structural dimensions (see Table 2). This trend is most pronounced in the compression ratio, though pairwise scores on sampled data reveal some inconsisten-

cies. Across pairwise metrics—ROUGE-L (lexical focus), BERTScore (semantic focus), and Tree Kernel distance (syntactic focus)—SCOUT, which includes natural human-robot dialogue, demonstrates significantly higher diversity compared to any VLA dataset. Human-written instructions from OASST2 also show considerably more diversity. For ALFRED, pairwise scores are similar to those of the more diverse VLA datasets, despite stronger results in intrinsic dimensionality and compression ratio. LLM-generated instruction-tuning datasets, particularly LLaVA-Instruct, also display limited diversity, although most metrics still report higher diversity than VLA datasets.

Analysis 1: Lexical Diversity. VLA datasets, particularly those with templated language, exhibit exceptionally low word counts, indicating poor lexical diversity (see Table 2). The robotics datasets contain significantly fewer unique words than instruction-tuning datasets. For instance, RT-1, TacoPlay, and LIBERO are particularly limited in this respect—RT-1 contains only 49 unique words. In contrast, the Bridge dataset showcases the highest number of unique words among the VLA datasets, even surpassing Language Table, which includes ten times more unique commands. Non-VLA datasets like ALFRED and SCOUT also contain more unique words than most VLA datasets, despite SCOUT having fewer commands than Bridge or Language Table.

This difference with instruction-tuning datasets is somewhat expected, as instruction-following datasets are not constrained by a physical environment and can reference a broader range of objects and actions. However, some variation—such as through synonyms or paraphrasing—might be expected even in robotics contexts. Yet, many VLA datasets show little to no lexical diversity.

For example, the full set of unique words in RT-1 is limited to the following words

> *bottle, apple, chip, upright, green, and, pick, chocolate, open, bag, 7up, over, blueberry, shelf, rxbar, bottom, redbull, door, paper, drawer, counter, brown, knock, on, plastic, bowl, move, left, coke, fridge, blue, jalapeno, close, right, sponge, into, place, orange, pepsi, water, from, white, of, rice, top, can, near, middle, banana.*

We also look into the lexical overlap between datasets. We analyze how much vocabulary is shared across datasets along the following POS categories: verbs, nouns, and adverbs (see Figure 7 in Appendix C.2). Nouns overall are the most widely shared category, likely because many robotic tasks involve similar objects (e.g., boxes, cans, drawers). Verbs are also shared, though to a lesser extent, likely constrained by the specific capabilities of each robot embodiment. Only four words appear in all datasets: move, close, open, and pick, which are the action verbs for the majority of tasks the robots are learning to perform.

Analysis 2: Semantic Diversity. Compared to other robotics and instruction-tuning datasets, the difference in VLA datasets is noticeable—low intrinsic dimensionality, low word counts, and repetition point to narrow and repetitive coverage of objects, actions, and tasks. The intrinsic dimensionality analysis, measured by the number of PCA components required to explain 95% of cumulative variance, reinforces this pattern and shows that VLA datasets are the least diverse (see Table 2's USE column). All four tested encoders show similar results (see Table 5 in Appendix D). This measure is not determined by the number of unique commands. While the number of PCA components strongly correlates with the number of unique unigrams, its relationship with the number of unique commands is notably weaker (see Figure 8 in the Appendix D). That further highlights the similarity between individual commands.

Even with a limited number of unique words in the RT-1 dataset, certain verb-object combinations are very frequent, such as "pick" and "banana," whereas some objects, like "white bowl" and "paper bowl" are rarely represented at all. However, other actions involving these words occur much less frequently, which could introduce potential biases (see Figure 3). Across datasets most objects co-occur with fewer than ten verbs, indicating limited task diversity (see Figures 22 and 21 in Appendix F). However, ALFRED and Language Table exhibit more balanced and varied distributions. While some constraints stem from limitations in manipulation capabilities, others appear artificial; for example, TacoPlay's stacked blocks could support richer interactions (e.g., "observe" or "tip"). For navigation datasets like SCOUT, we examine the diversity of adverbials, which modify actions in ways that convey nuance in direction, location, manner, and time. As shown in Figure 17 the lexical field is heavily skewed toward directional adverbs—notably terms such as "north," "forward,"

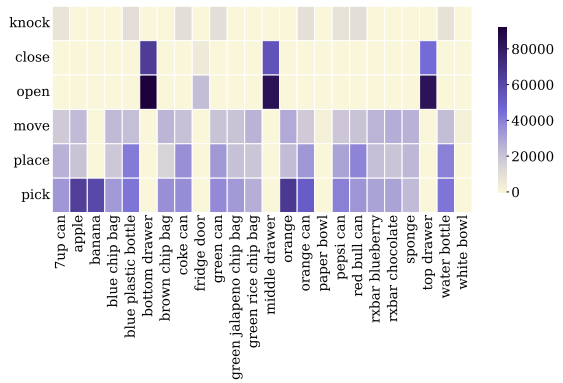


Figure 3: **Analysis 2: Semantic Diversity** Verb and direct object co-occurrence frequencies in the manually annotated RT-1 dataset. The heatmap highlights limited verb diversity across plausible actions; for instance, *banana* is frequently "picked" but never "moved." The rare verb *knock* appears mostly with can-shaped objects, despite being equally applicable to others like an upright sponge.

"south," and related spatial indicators. These directionals form the backbone of navigational grounding in SCOUT; however, greater emphasis on how the robot moves, e.g., "fast" or "slow" could be beneficial.

Analysis 3: Structural Diversity. Multi-step instructions are frequent in robotics datasets, whereas negations, conditionals, and cyclical structures are almost entirely absent (see Figure 5 for distributions and Table 10 in the Appendix G for examples). Multi-step commands are the most prevalent across all datasets, reflecting a strong bias toward procedural, linear task decomposition—particularly notable in LIBERO. In contrast, datasets like RT-1 and SCOUT contain fewer multi-step instructions and favor shorter, atomic actions. Negation and conditional constructions occur in fewer than 2% of cases, suggesting that many benchmarks fail to capture logical disjunctions, exception handling, or constraint-driven behaviors—elements essential for safe and flexible deployment. Cyclical or loop-like structures, which are common in real-world tasks, are similarly underrepresented, with only SCOUT and ALFRED showing a modest signal. Overall, these patterns point to a structural bias in current datasets toward flat, step-by-step formulations, with limited support for complex task logic.

Analyzing part-of-speech (POS) patterns yields similar insights. This analysis examines the grammatical structure of commands, specifically focusing on how words are arranged using POS patterns. The most frequent patterns typically start with a verb and describe the object using either a single noun, an adjective followed by a noun, or two nouns, followed by an adposition and a new object description. In the cases of RT-1 and TacoPlay, the most frequent pattern constitutes 11% and 24% of all patterns, respectively (see Figure 4). This reliance on repetitive structures may make it harder for models to generalize to more complex instruc-

RT-1 (Brohan et al., 2023)

VERB→ NOUN→ NOUN→ ADP→ ADJ→ NOUN (11%)
place water bottle into white bowl
VERB→ NOUN→ NOUN→ ADP→ NOUN→ NOUN (7%)
move redbull can near 7up can

BRIDGE (Ebert et al., 2022)

VERB→ DET → NOUN→ ADP→ DET → NOUN→ PUNCT (3%)
Place the mushroom behind the spatula.
VERB→ DET → NOUN→ ADP→ DET → NOUN→ ADP→ DET → NOUN→ PUNCT (3%)
Move the spoon to the left of the napkin.

TacoPlay (Rosete-Beas et al., 2022)

VERB→ DET → ADJ→ NOUN→ ADP→ DET → NOUN (24%)
put the purple block on the table
VERB→ DET → ADJ→ NOUN→ ADP→ DET → ADJ→ NOUN (6%)
put the pink object inside the left cabinet

Language Table (Lynch et al., 2023)

VERB→ DET → ADJ→ NOUN→ ADV→ ADP→ DET → ADJ→ NOUN (4%)
move the blue cube right to the yellow hexagon
VERB→ DET → ADJ→ NOUN→ ADP→ DET → ADJ→ NOUN (4%)
push the green circle towards the green star

Figure 4: **Analysis 3: Structural Diversity** Most frequent POS patterns per dataset on unique sentences and their relative frequency and example.

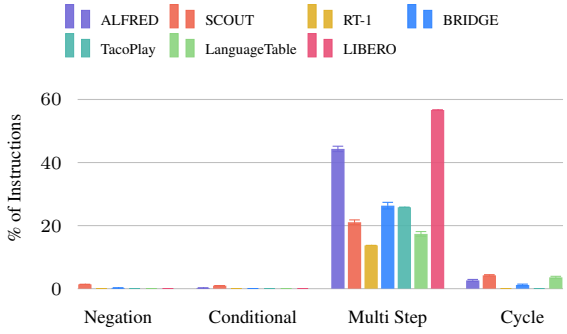


Figure 5: **Analysis 3: Structural Diversity** Percentage of instructions exhibiting four structural phenomena: negation, conditionality, multi-step sequencing, and cyclic repetition. For datasets containing fewer than 600 unique sentences, annotations were performed manually. For those with more than 600 unique sentences, annotations were generated using an automated pipeline. Standard error bars reflect labeling uncertainty estimated from a manually reviewed subset of 500 randomly sampled commands per dataset.

tions. Refer to Figures 11 and 12 in Appendix E for more qualitative examples of dominant patterns. Figure 13 offers an aggregated view across datasets, and Figures 14 and 15 group by dataset.

The command length distribution across seven datasets reveals a preference for short commands that fall within the range of 3 to 15 words. This further highlights the dominance of concise phrasing, which may limit exposure to more complex linguistic structures, e.g., multi-clause, multi-step instructions.

6 Conclusion & Future Directions

In this work, we characterize linguistic diversity in widely used VLA datasets using complementary lexical, semantic, and structural metrics. Across VLA datasets, the language is highly repetitive: there are few unique commands, limited lexical variation, and narrow coverage of objects, actions, and tasks. Syntactic diversity is also constrained: multi-step commands are common, while negation and conditionals are rare.

Prior work suggests that linguistic diversity may matter for generalization. Our analyses provide a starting point for doing so in VLA datasets by identifying which aspects of language are currently underrepresented. We can leverage these insights to improve linguistic diversity using several possible strategies: (i) **Targeted augmentation**—using paraphrasing and template-based generation to expand underrepresented patterns; (ii) **Cross-domain transfer**—selectively leveraging linguistically richer instructional corpora (e.g., procedural text, situated dialogue) to complement VLA data; and (iii) **Annotation guidance**—encouraging data collection that explicitly includes missing constructions. Among these, our analyses most directly support both (i) targeted augmentation and (iii) annotation practices.

Targeted augmentation enables us to improve existing datasets' linguistic diversity without building new datasets. Based on our analyses, we can leverage syntactic similarity metrics—such as TreeKernel methods or POS pattern histograms (Analysis 3)—to guide LLMs in generating paraphrases through phrase permutation or inversion. In many cases, the vocabulary within current datasets is limited (Analysis 1). Here, LLM-guided synonym replacement could help increase lexical diversity.

A more resource-intensive—but higher-quality—approach is to use these findings to guide future data collection. For example, Analysis 1 reveals strong n-gram overlap and frequent word duplication. To mitigate this, data collection interfaces could prompt users in real time to rephrase instructions using alternative wording or synonyms. Analysis 2 shows strong co-occurrence patterns between specific verbs and nouns in the RT-1 dataset, suggesting the presence of potentially superficial correlations. Addressing this may require encouraging more varied verb–object pairings during data collection. Finally, Analysis 3 highlights that POS patterns are highly clustered within datasets and that structural diversity is dominated by multi-step commands. To counteract this, participants could be prompted to use more descriptive and varied language—for example, by incorporating adverbial phrases—since the repetitive nature of dataset construction may otherwise discourage linguistic richness. Our analysis also suggests that more interactive data collection methods—such as SCOUT's Wizard-of-Oz approach—can yield greater lexical diversity (Analysis 1) and richer structural variation (Analysis 3). This points to a promising pathway for generating linguistically diverse data.

7 Limitations

Our analysis is motivated by the goal of enabling more generalist policies and stronger robotics foundation models, with language as a core modality. The patterns we identify may be less critical in narrower settings or application domains with constrained task distributions. Moreover, not all linguistic phenomena are equally relevant in all pipelines: for example, negation may be crucial for more atomic commands, whereas conditionals may matter less when planning is delegated to an external LLM.

We also acknowledge that object diversity is often limited by the costs of acquiring props and training novel tasks which biases robotics datasets toward low linguistic diversity. However, our work points toward the possibility that certain object categories are overrepresented–indicating that investment in new prop objects or alternative environment scenery (e.g., beyond kitchen environments) may be more beneficial.

Although robotics datasets are inherently multimodal, our study focuses exclusively on their textual components. Language functions in tan-

dem with visual input and action trajectories in embodied systems; because we do not evaluate cross-modal alignment, our findings do not capture potential inconsistencies between commands and the corresponding trajectories or visual scenes.

Individual metrics also have known limitations and may not capture linguistic diversity in full. To mitigate this, we report a diverse set of complementary lexical, semantic, and structural measures, and supplement them with manual analyses and multiple methodological perspectives where appropriate.

Finally, while our study covers several widely used VLA and robotics datasets that reflect dominant trends in the field, the results may not generalize to all instruction-following datasets in embodied AI.

Acknowledgments

This study was in part funded by the Estonian Research Council grants PRG3237 and PRG2006; and LDRD grant 20250048DR (U.S. DOE NNSA Contract 89233218CNA000001) (LA-UR-25-24853).

References

Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, and 33 others. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

Emily M. Bender. 2019. The benderrule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, and 5 others. 2024. π_0: A vision-language-action flow model for general robot control. *Preprint*, arXiv:2410.24164.

Anthony Brohan and 1 others. 2023. Rt-1: Robotics transformer for real-world control at scale. *Preprint*, arXiv:2212.06817.

Daniel Cer and 1 others. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. 2025. Training strategies for efficient embodied reasoning. *Preprint*, arXiv:2505.08243.

Embodiment Collaboration and 1 others. 2024. Open x-embodiment: Robotic learning datasets and rt-x models. *Preprint*, arXiv:2310.08864.

Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Madry, Andrew Ilyas, and Roberto Martín-Martín. 2025. Datamil: Selecting data for robot imitation learning with datamodels. *Preprint*, arXiv:2505.09603.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mathieu Dehouck and Carlos Gómez-Rodríguez. 2020. Data augmentation via subtree swapping for dependency parsing of low-resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3818–3830, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Danny Driess, Jost Tobias Springenberg, brian ichter, LILI YU, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. 2025. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Paul-Ambroise Duquenne and 1 others. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint*.

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. 2022. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems*.

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. 2010. Intrinsic dimension estimation of data by principal component analysis. *Preprint*, arXiv:1002.2050.

Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. 2025. A taxonomy for evaluating generalist robot policies. *Preprint*, arXiv:2503.01238.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012.

Shresth Grover, Akshay Gopalkrishnan, Bo Ai, Henrik I. Christensen, Hao Su, and Xuanlin Li. 2025. Enhancing generalization in vision-language-action models by preserving pretrained representations. *Preprint*, arXiv:2509.11417.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Joey Hejna, Chethan Anand Bhateja, Yichen Jiang, Karl Pertsch, and Dorsa Sadigh. 2025. Remix: Optimizing data mixtures for large scale imitation learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 145–164. PMLR.

Matthew Honnibal and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, and 17 others. 2025. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *Preprint*, arXiv:2504.16054.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model. *Preprint*, arXiv:2406.09246.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2025. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023a. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Preprint*, arXiv:2306.03310.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Stephanie M. Lukin and 1 others. 2024. SCOUT: A situated and multi-modal human-robot dialogue corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458, Torino, Italia. ELRA and ICCL.

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. 2023. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, pages 1–8.

Yash Mahajan, Naman Bansal, Eduardo Blanco, and Santu Karmaker. 2024. ALIGN-SIM: A task-free test bed for evaluating and interpreting sentence embeddings through semantic similarity alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7393–7428, Miami, Florida, USA. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–120, Trento, Italy. Association for Computational Linguistics.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Kishore Papineni and 1 others. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Fast: Efficient action tokenization for vision-language-action models. *Preprint*, arXiv:2501.09747.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint*.

Nils Reimers and 1 others. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Erick Rosete-Beas and 1 others. 2022. Latent plans for task agnostic offline reinforcement learning.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.

Haoyue Shi, Karen Livescu, and Kevin Gimpel. 2021. Substructure substitution: Structured data augmentation for nlp. *Preprint*, arXiv:2101.00411.

Mohit Shridhar and 1 others. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Fumiya Uchiyama, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Which programming language and what features at pre-training stage affect downstream logical inference performance? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18139–18149, Miami, Florida, USA. Association for Computational Linguistics.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Michel Verleysen and John A. Lee. 2013. Nonlinear dimensionality reduction for visualization. In *Neural Information Processing*, pages 617–622, Berlin, Heidelberg. Springer Berlin Heidelberg.

Homer Walke and 1 others. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*.

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. 2024. Ladev: A language-driven testing and evaluation platform for vision-language-action models in robotic manipulation. *Preprint*, arXiv:2410.05191.

Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2025. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. In *9th Annual Conference on Robot Learning*.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2024. Robotic control via embodied chain-of-thought reasoning. In *8th Annual Conference on Robot Learning*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yan Zhang, Zhaopeng Feng, Zhiyang Teng, Zuozhu Liu, and Haizhou Li. 2023. How well do text embedding models understand syntax? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9717–9728, Singapore. Association for Computational Linguistics.

Yubo Zhang and 1 others. 2020. Diagnosing the environment bias in vision-and-language navigation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 890–897. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. 2025. NegVQA: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716, Vienna, Austria. Association for Computational Linguistics.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: Measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60644–60673. PMLR.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Ran Cheng, Yaxin Peng, Chaomin Shen, and Feifei Feng. 2025. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. *Preprint*, arXiv:2502.14420.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Qualitative Features of EAI datasets

We conducted an informal qualitative review of the examined datasets and highlighted interesting attributes, summarized in Table 3.

On Conversational Strengths. The SCOUT dataset exhibits a distinct dialogue structure that differentiates it from traditional instruction-following datasets. Rather than adhering to a rigid, directive style, its dialogues often involve an exploratory or inquiry-based approach, as seen in exchanges like "move west uh zero point five meters" and "...and then the last question here anything that indicates the environment was recently occupied". This interactive nature may offer advantages for EAI by allowing more adaptive responses. For example, in cases where instructions involve complex spatial reasoning; e.g., placing an object in a specific but ambiguous location, the dataset's conversational format could aid in disambiguation.

On Cultural Knowledge. One of the more striking aspects of the BRIDGE dataset is its incorporation of multicultural culinary terminology, despite being primarily monolingual (English). Unlike many Western-centric datasets, BRIDGE includes references to diverse cooking utensils and ingredients, such as purkoli (broccoli), brinjal (eggplant), brezzela (eggplant), capsicum (bell pepper), quince fruit, nigiri, wok, and kadai. This linguistic diversity suggests a broader representation of cultural knowledge, making incremental progress toward addressing concerns raised in prior work on dataset biases (Bender et al., 2021; Bender, 2019). Specifically, it challenges the tendency for data collection to reflect primarily Western, white audiences. Additionally, BRIDGE captures subtle social characteristics of human perception, such as humor, evidenced by an annotation that describes a mushroom toy as a "phallic looking item."

On "Common Sense" Reasoning. A recurring challenge across real-world datasets is the disconnect between world knowledge, common-sense reasoning, and practical instruction execution. While

Theme	Example Instruction(s)
Cultural Terms (BRIDGE)	"put the kadai on the stove", "grab the brinjal from the drawer"
Unsafe Action (ALFRED)	"store a knife in a microwave", "stab the tip of the knife into the table"
Commonsense Violation (ALFRED)	"Put an egg in a pan in the fridge"
Commonsense Violation (BRIDGE)	"take sushi out of the pan"

Table 3: Selected examples illustrating conversational structure, cultural variation, and commonsense inconsistencies across EAI datasets.

BRIDGE and ALFRED aim to ground tasks in realistic environments, many instructions contain fundamental inconsistencies or implausible directives. In ALFRED, for example, commands such as "open refrigerator, place potato to the right of tomato on second shelf of refrigerator, close refrigerator, open refrigerator, pick up potato from refrigerator, close refrigerator" expose rigid, mechanical assumptions about human behavior. Additionally, one must ask what has been accomplished by storing a potato in a refrigerator and then removing said potato in a matter of seconds. Another example from ALFRED includes, "Put an egg in a pan in the fridge." More concerning, and at times, unintentionally amusing, are instances of potentially unsafe or property-damaging instructions, such as "place a heated slice of tomato on a counter and **store a knife in a microwave**" or "**stab the tip of the knife into the wooden table**, in front of the gray plate closest to the lettuce." While a robot damaging a kitchen table may be preferable to microwaving a knife, these examples highlight inconsistencies in world knowledge modeling within these datasets. Similar anomalies appear in BRIDGE, where commands such as "take sushi out of the pan," "put sushi in pot...," and "put spatula in pan" suggest an oversimplified understanding of object affordances, human behavior, and broader world and cultural knowledge. If the broader EAI community sees embodiment as a necessary step toward elevating the representational learning of single-modality models, e.g., LLMs, we ought to discourage dataset collectors from building illogical "common-sens"" associations.

B Text Preprocessing

For calculating the metrics in Table 2, we preprocess the text by splitting the text into sentences, standardizing white space, and removing punctuation. We calculate the statistics using the combination of `spacy` (Honnibal et al., 2020) and `pandas` (pandas development team, 2020) methods.

To measure unique words, we removed punctuation from our cleaned text. Then, we concatenated all sentences in a particular corpus into a single string. We tokenized the text into unigrams using Python's `.split()`, converted the resulting list into a `.set()` to obtain unique tokens, and computed the number of unique words as the size of that set.

For all experiments, we cleaned the SCOUT (Lukin et al., 2024) dataset of user role tags and tags that indicate filler words, e.g., "um", silence, and noise. Due to the complexity of this data, we focus our initial analysis only on the "robot commander" dialogue, with plans to expand our analysis to all roles in the future and to incorporate filler filtering in the text cleaning pipeline.

C Lexical Diversity Extended Results

In addition to the metrics discussed in the main body, we measured sentence length (Section C.1), lexical overlap (Section C.2), and some additional lexical diversity metrics (Section C.3).

C.1 Sentence Length

The majority of commands contain fewer than ten words (see Figure 6). Command lengths are capped at a maximum of 30 words for display purposes. OASST shows wide standard deviation because of outlier examples from the instruction datasets that contain up to 100+ words. Despite the impossibility of a negative sentence length we display the standard deviation exactly as other datasets. The *Instruction-Tuning and NLU Datasets* are all shown in grayscale with various hatching to distinguish. Generally, these datasets share similar sentence lengths as Language Table, LIBERO, ALFRED, however greatly outperform the aforementioned robotics datasets against our diversity measures (c.f. Tables 2 and 5).

C.2 Lexical Overlap

To assess how much vocabulary is shared across datasets, we examine the distribution of words across three part-of-speech (POS) categories:

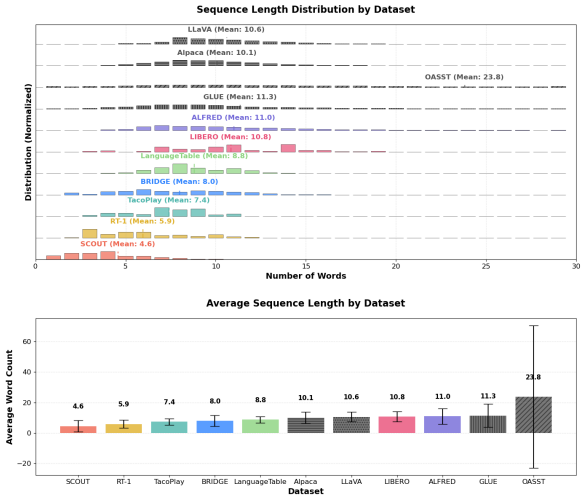


Figure 6: **Analysis 1: Lexical Diversity** Distribution of command lengths across six examined EAI datasets.

nouns, verbs, and adverbs. We use dependency parsing to extract tokens by their POS tags. We then construct a dataset–word matrix that records how often each word appears in more than one dataset. This allows us to visualize lexical overlap using a heatmap (Figure 7).

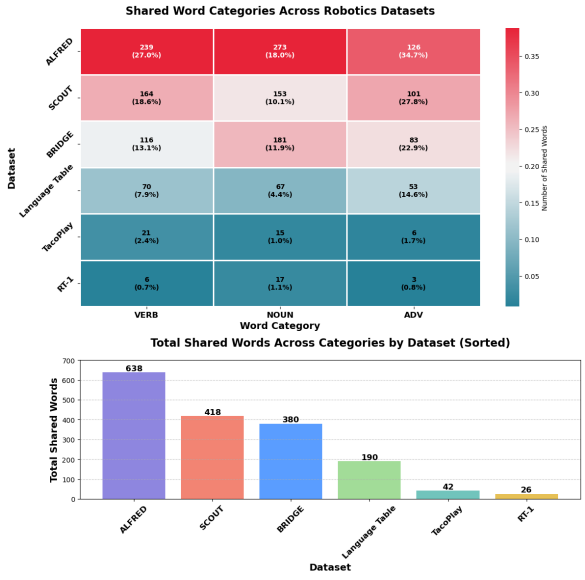


Figure 7: **Analysis 1: Lexical Diversity** Shared POS categories across datasets. Using ALFRED as a pretraining dataset is advantageous because it has the greatest amount of lexical coverage across the examined EAI datasets.

C.3 Additional Metrics: Levenshtein, Jaccard, BLEU-4

In addition to metrics shown in 2, we measured Levenshtein, Jaccard, and, following previous work (Zhang et al., 2020), BLEU-4. Given that these methods entail pair-wise comparisons, we perform 1,000 commands to obtain these scores across 3 trials, as with other pair-wise metrics. The results are in Table 4, and the scores agree with other metrics.

D Intrinsic Dimensionality Analysis

A notable limitation of our methodology is using linear dimensionality reduction techniques, specifically PCA, to assess data that may lie on a nonlinear manifold, as is often the case with LLM-encoded datasets. While PCA assumes linearity, this limitation does not significantly undermine our analysis. In fact, it likely results in an *overestimation* of the intrinsic dimensionality, since PCA cannot exploit underlying nonlinear relationships in the data (Verleysen and Lee, 2013). For our purposes, this effect only further underscores the discrepancy between the structure of robotics datasets and the more diverse language representations found in other datasets we studied.

The full results with four different sentence encoders are in Table 5. We also measured the correlation between the number of PCA components required to explain 95% variance and language statistics across EAI datasets (see Figure 8).

Although the conclusions of this analysis are reinforced by our more interpretable feature-based methods (see Table 2); in future work, we would like to strengthen this effort.

E POS Patterns

We implemented a large-scale dependency parsing pipeline using an LLM to extract POS and dependency parse patterns, leveraging multi-GPU parallel processing for efficiency. Each GPU independently processed a subset of instructions using `DeepSeek-R1-Distill-Qwen-32B` (DeepSeek-AI et al., 2025), a state-of-the-art instruction-following LLM. The model was loaded in 8-bit quantized format to optimize memory usage, and batch $b = 10$ processing was employed to maximize throughput. The prompts for the model followed a structured format (see Figure 9), instructing it to perform dependency parsing and return results in valid JSON format. The output JSON included:

- The original instruction
- A tokenized breakdown, where each word was annotated with its:

Dataset	Levenshtein ↑	Jaccard ↓	BLEU-4 ↓
Instruction-Tuning Datasets			
OASST2 (Köpf et al., 2023)	**78.238 ± 1.789**	**0.033 ± 0.001**	**0.0001 ± 0.0001**
Alpaca (Taori et al., 2023)	51.382 ± 0.537	0.061 ± 0.001	0.0003 ± 0.0001
LLaVA-Instruct (Liu et al., 2023b)	46.465 ± 0.223	0.130 ± 0.002	0.002 ± 0.001
Language-Focused Robotics Datasets			
ALFRED (Shridhar et al., 2020)	**46.695 ± 0.883**	0.128 ± 0.004	0.003 ± 0.000
SCOUT (Lukin et al., 2024)	24.512 ± 0.946	**0.052 ± 0.002**	**0.002 ± 0.001**
VLA Datasets			
RT-1 (Brohan et al., 2023)	28.143 ± 0.413	0.138 ± 0.001	0.026 ± 0.006
BRIDGE (Walke et al., 2023)	**35.139 ± 0.180**	**0.088 ± 0.004**	**0.003 ± 0.000**
TacoPlay (Rosete-Beas et al., 2022)	27.705 ± 0.137	0.188 ± 0.003	0.020 ± 0.001
Language Table (Lynch et al., 2023)	32.206 ± 0.171	0.198 ± 0.002	0.010 ± 0.001
LIBERO (Liu et al., 2023a)	34.269 ± 0.188	0.248 ± 0.006	0.064 ± 0.003

Table 4: **Analysis 1: Lexical Diversity** Subset of text similarity measures: Levenshtein distance, Jaccard similarity, and BLEU-4. Arrows indicate that higher Levenshtein and lower Jaccard/BLEU-4 correspond to greater diversity.

Dataset	SBERT ↑	USE ↑	SONAR ↑	CLIP ↑
Instruction-Tuning				
OASST2 (Köpf et al., 2023)	**396**	254	**754**	361
Alpaca (Taori et al., 2023)	350	231	637	338
LLaVA-Instruct (Liu et al., 2023b)	245	184	540	279
Language-Focused Robotics Datasets				
ALFRED (Shridhar et al., 2020)	165	**159**	**406**	**198**
SCOUT (Lukin et al., 2024)	**194**	148	295	181
VLA Datasets				
RT-1 (Brohan et al., 2023)	27	33	42	35
BRIDGE (Walke et al., 2023)	**115**	**125**	**239**	**149**
TacoPlay (Rosete-Beas et al., 2022)	31	42	41	36
Language Table (Lynch et al., 2023)	57	86	108	71
LIBERO (Liu et al., 2023a)	32	34	44	33

Table 5: **Analysis 2: Semantic Diversity** The minimum number of PCA components to explain 95% variance for each dataset. A greater number of components represents stronger diversity.

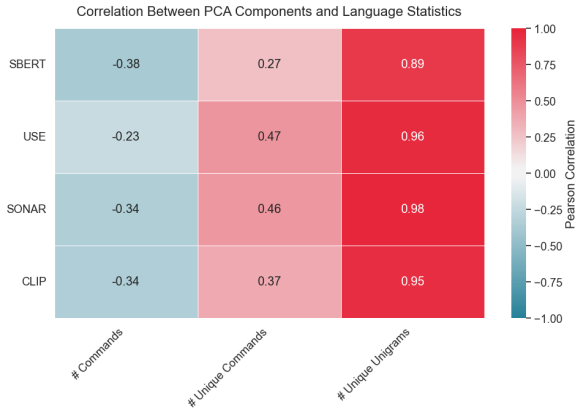


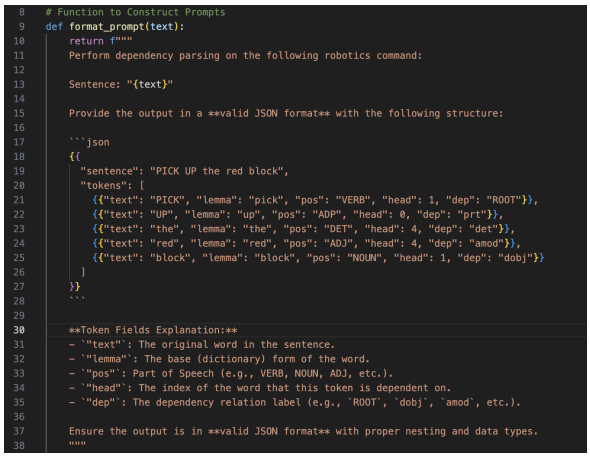
Figure 8: Correlation between the number of PCA components required to explain 95% variance and language statistics across EAI datasets. PCA components derived from SBERT, USE, SONAR, and CLIP embeddings are compared against the number of commands, unique commands, and unique unigrams in each dataset. Strong positive correlations are observed between unique unigrams and all embedding models, particularly SONAR and USE. In contrast, the total number of commands shows weak or negative correlation with embedding diversity

- Lemma (root form)
- Part of speech (POS) tag
- Syntactic head (parent word in the dependency tree)
- Dependency label (e.g., ROOT, direct object, modifier, etc.)

For qualitative examples related to each POS pattern, please refer to Figures 11 and 12.

The decision to use LLMs for the POS tagging task was driven by an exploratory, qualitative review of preliminary outputs from several traditional NLP tools. We evaluated spaCy models, including `en_core_web_sm` and `en_core_web_trf`, as well as StanfordNLP's Stanza POS tagging model (`POSProcessor`). Across these models, including transformer-based variants, we observed recurring difficulties with the unconventional object naming conventions in the RT-1 dataset (e.g., `rxbar`). Several models also struggled to reliably distinguish between the noun "can" (e.g., referring to 7up cans in the dataset) and the verb "can" (expressing ability). During informal testing, the DeepSeek models appeared to handle these cases better more consistently; so we opted to build our POS tagging workflow around them. However, to emphasize, we did not conduct a quantitative comparison across parsers.

The **BRIDGE** dataset is heavily characterized by prepositional phrases, frequently structuring instructions that specify spatial relationships between



```
# Function to Construct Prompts
def format_prompt(text):
    return f"""
    Perform dependency parsing on the following robotics command:

    Sentence: "{text}"

    Provide the output in a **valid JSON format** with the following structure:

    ```json
    {{
      "sentence": "PICK UP the red block",
      "tokens": [
        {{"text": "PICK", "lemma": "pick", "pos": "VERB", "head": 1, "dep": "ROOT"}},
        {{"text": "UP", "lemma": "up", "pos": "ADP", "head": 0, "dep": "prt"}},
        {{"text": "the", "lemma": "the", "pos": "DET", "head": 4, "dep": "det"}},
        {{"text": "red", "lemma": "red", "pos": "ADJ", "head": 4, "dep": "amod"}},
        {{"text": "block", "lemma": "block", "pos": "NOUN", "head": 1, "dep": "dobj"}}
      ]
    }}
    ```

    Token Fields Explanation:
    - `"text"`: The original word in the sentence.
    - `"lemma"`: The base (dictionary) form of the word.
    - `"pos"`: Part of Speech (e.g., VERB, NOUN, ADJ, etc.).
    - `"head"`: The index of the word that this token is dependent on.
    - `"dep"`: The dependency relation label (e.g., `ROOT`, `dobj`, `amod`, etc.).

    Ensure the output is in **valid JSON format** with proper nesting and data types.
    """
```

Figure 9: Prompt used in dependency parse work.

objects and the environment. This results in a high frequency of ADP (adpositions), NOUN (nouns), and DET (determiners), forming patterns, e.g. "put the spoon on the cloth", "put the mangoes in a pan", and "Move the spatula near the egg." While this structure ensures precision in command execution, it lacks syntactic variation beyond simple prepositional constructs, potentially limiting generalization to more complex spatial reasoning tasks.

RT-1, in particular, exhibits highly repetitive syntactic patterns, as seen in commands like "place 7up can into middle drawer," "place water bottle into white bowl," and "place rxbar blueberry into bottom drawer." Similarly, TacoPlay demonstrates significant syntactic redundancy, with instructions such as "place the purple block on the table," "store the pink object in the drawer," and "slide the yellow block to the right." This lack of linguistic variability, likely due to the template-driven generation of these datasets, may limit a model's ability to generalize to more complex instructions, particularly those involving hierarchical dependencies or compound actions.

SCOUT introduces more numerical expressions and adverbial structures, implying an instructional style where robots may be required to count, measure, or modify behaviors dynamically, e.g., "move south four feet", "turn right twenty degrees", "go forward one meter". However, its emphasis on concise command structures might underrepresent more complex multi-step directives.

The POS histograms in Figures 13-15 reveal a long-tailed distribution in TacoPlay, SCOUT, and RT-1, where the frequency of syntactic structures drops sharply after the first or second most common parse pattern. Such patterns indicate a reliance

Dataset	Standard Error Rate
TacoPlay	0.01 ± 0.01
RT-1	0.00 ± 0.02
SCOUT	0.01 ± 0.014
ALFRED	0.01 ± 0.01
BRIDGE	0.02 ± 0.02
LanguageTable	0.01 ± 0.01
LIBERO	0.01 (full dataset)

Table 6: Standard error rates across datasets. Values are reported as mean $\pm$ 95% confidence interval with $n = 200$, unless otherwise noted.

on repetitive syntactic templates, which may limit a model's ability to generalize to linguistically varied instructions. Language Table shows the longest and most evenly distributed bar set among all datasets, with no single POS pattern dominating. Language Table sets the upper bound for linguistic diversity among embodied AI datasets and should be more widely used. However, for datasets like RT-1, we recommend that synthetic data augmentation could help mitigate this imbalance by introducing greater syntactic variability, such as tree-based reordering techniques, inspired by data augmentation in machine translation (Dehouck and Gómez-Rodríguez, 2020; Shi et al., 2021), could be adapted to generate syntactic variants of robotic commands while preserving their semantics.

On Annotation Quality. Table 6 presents standard error rates on POS tagging using `GPT-5.2` as a reference baseline. No human intervention (i.e. additional post processing) was applied to these outputs.

F Verb, Direct Object, Adverbial Diversity.

To extract verb, direct object, and adverbial features, an author manually annotated the LIBERO, TacoPlay, RT-1, SCOUT datasets, and Bridge datasets. For larger datasets: Language Table and ALFRED we implemented a large-scale annotation pipeline using `R1-Distill-Qwen-14B` (DeepSeek-AI et al., 2025). The model was loaded in 8-bit quantized format to optimize memory usage, and batch $b = 10$ processing was employed to maximize throughput. The prompts for the model followed the format shown in Figure 10. We implemented in-context learning (ICL) on Language Table to enhance accuracy by retrieving sentence-specific examples using TF-IDF similarity. Despite using LLMs, all annotations were manually reviewed to ensure consistency, including lemma-

tizing verbs, removing duplicates, and normalizing synonymous expressions (e.g., “pick" vs. “pick up"). This hybrid method enabled the construction of high-quality annotations for downstream analysis. Results are provided in Figures 22, 21, and 17.

On Object and Adverbial Diversity. We assessed how many distinct verbs are used with each direct object for manipulation datasets. Low counts suggest limited interaction diversity, sometimes due to real-world constraints, but often due to overly templated instruction generation. Direct object structures are less relevant for navigation-focused datasets, instead how an instruction is followed, e.g., directional terms (e.g., “north," “forward"), location-based modifiers (e.g., “around," “inside"), manner descriptors (e.g., “slowly," “directly") are more relevant.

On Numeric Generalization. As VLA models are increasingly expected to interpret numerical quantities (e.g., distances, angles) in an end-to-end manner, the distribution of numerical values in navigation instructions becomes more critical. Figure 18 shows that numbers like “two," “three," and “five" are relatively common in SCOUT, while values such as “seven," “eight," or “twelve" are rare. ALFRED (see Figure 19) appears more tailed and its numeric coverage is weaker than SCOUT; however, the overall representation of numerics is greater due to dataset size. This sparsity raises concerns about whether models trained on these datasets can interpolate or generalize to underrepresented numerical instructions. For example, can a robot correctly interpret “move seven meters" if it has never encountered that number in training? What if it has only encountered meters but is given a command in yards? What if the command contains common shortcuts, such as using 4K to refer to 4,000? Future research should investigate the impact of numeric and unit sparsity on navigation performance and explore methods for balancing numerical distributions during data collection or augmentation.

On Annotation Quality. For the dependency parsing results focused on direct object and adverbial analysis, we employed a multi-stage human-in-the-loop annotation pipeline. Initial verb–direct object pair predictions follow from our aforementioned procedure, as a coarse draft. This first-stage output was evaluated in a binary “good”/“bad” manner based on whether the correct verb–object pairs were present or if extra incorrect pairs were included. These predictions were not used as final outputs. The annotations underwent 2-3 rounds of cleaning, standardization, and corrections before being reported. We provide quantitative details on the initial draft error rates in Table 7, but we emphasize that these first-round predictions were not used in generating the final results.

G Instruction Structure Analysis

To analyze the compositional structure of language in robotics datasets, we use LLM-generated feature information (see Appendices E and F) to construct heuristics for detecting four types of instruction-level patterns: negation, conditionality, multi-step sequencing, and cyclical structures. These patterns are identified through string-matching techniques and syntactic cues extracted from dependency parses and part-of-speech tags.

- **Negation** was detected using syntactic cues like neg dependencies and lexical markers (e.g., “not”, “don’t”, “never”).
- **Conditionality** was identified via subordinating conjunctions (e.g., “if”, “unless”) and dependency markers indicating conditional clauses.
- **Multi-step** sequencing was inferred from coordinating conjunctions (e.g., “and”, “then”), punctuation, or imperative chaining.
- **Cyclical** patterns were identified using repeat verbs (“again”, “repeat”) or constructions indicating iteration or loops.

For each instruction, we annotated binary indicators for each structure type and aggregated them to compute relative frequencies across datasets. Quantitative results are presented in Figure 5, and representative examples are shown in Table 10. These results help reveal structural tendencies in instruction design; particularly, the dominance of linear, stepwise instruction formats and the underrepresentation of more complex, logic-driven patterns.

On Annotation Quality. In Figure 5, for datasets with fewer than 600 unique instructions, we manually annotated all examples and computed inter-annotator agreement (IAA) between two annotators. Human–human annotations consisted of binary judgments (“yes”/“no”) indicating whether the LLM-generated annotation was correct, independent of the specific structural category predicted. Inter-annotator agreement was

Dataset	Stage 1 LLM Standard Error Rate	Final Stage
TacoPlay	0.34 (full dataset)	Manually annotated full dataset
RT-1	0.23 (full dataset)	Manually annotated full dataset
SCOUT	–	Manually annotated full dataset
ALFRED	0.07 ± 0.00 (n=101, 95% CI)	Manually annotated full dataset
BRIDGE	0.28 ± 0.04 (n=590, 95% CI)	Manually annotated full dataset
LanguageTable	0.08 ± 0.04 (n=200, 95% CI)	Manually annotated full dataset
LIBERO	0.26 ± 0.09 (n=101, 95% CI)	Manually annotated full dataset

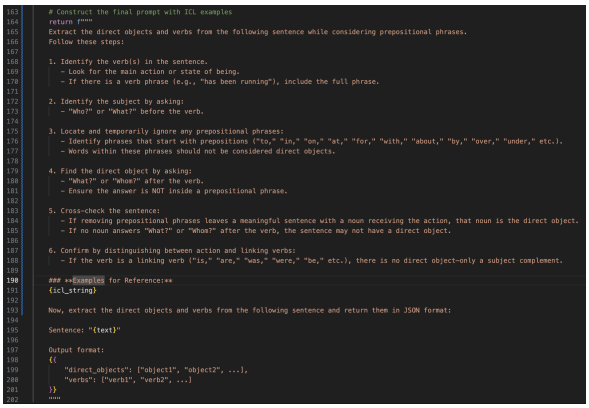
Table 7: Stage 1 LLM standard error rates and final annotation stage across datasets. Error rates are reported either on the full dataset or as mean $\pm$ confidence interval (95%) with sample size n.

Dataset	Negation	Conditional	Multi Step	Cycle
Language-Focused Robotics Datasets				
ALFRED (Shridhar et al., 2020)	22	275	56026	3313
SCOUT (Lukin et al., 2024)	122	85	1890	379
VLA Datasets				
RT-1 (Brohan et al., 2023)	0	0	82	0
BRIDGE (Walke et al., 2023)	27	2	3113	139
TacoPlay (Rosete-Beas et al., 2022)	0	0	104	0
Language Table (Lynch et al., 2023)	26	6	22164	4579
LIBERO (Liu et al., 2023a)	0	0	959	0

Table 8: **Analysis 3: Structural Diversity** Further details on the Instruction Structure Analysis. Raw counts corresponding to Figure 5.

then computed to assess consistency in these correctness judgments. For the smaller datasets (RT-1, TacoPlay, and LIBERO), annotators achieved perfect agreement, with Cohen's $\kappa = 1.0$. This may be due to the templated construction of the instructions for these datasets.

For datasets with more than 600 unique instructions, we randomly sampled 500 unique commands for manual annotation. We computed human agreement on this subset and used the same subset to estimate standard errors. To compute LLM–human agreement (IAA H-L) and the associated standard errors (SE_CATEGORY), one annotator additionally recorded category-specific correctness (see Table 9). The low agreement in multi-step was due to a systematic ambiguity regarding whether compound objects (e.g., 'move X and Y') should be labeled by their syntax (single-step) or their robotic execution (multi-step).



(a) Verb–direct object prompt example used in the **"Verb, Direct Object, Adverbial Diversity"** section.

```
    # Format in-context learning examples
    icl_string = "\n".join(
        f"""Example {i+1}:
    Sentence: "{ex['example_sentence']}"
    Output:
    {{
        "direct_objects": {ex['direct_objects']},
        "verbs": {ex['verbs']}
    }}\n"""
        for i, ex in enumerate(icl_examples)
        if ex  # Ensuring valid examples are included
    )
```

(b) In context learning string generated by TF-IDF distance k-nearest neighbors.

Figure 10: Prompts used in direct object and verb parsing tasks for instruction analysis.

Dataset	POS Pattern	Example Sentences
TacoPlay		
	VERB → DET → ADJ → NOUN → ADP → DET → NOUN	put the purple block on the table slide the purple block to the left place the yellow block on the table
	VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN	put the pink object inside the left cabinet put the yellow block inside the right cabinet place the purple block inside the right cabinet
	VERB → DET → ADJ → NOUN → CCONJ → VERB → PRON → ADV	take the purple block and rotate it right take the yellow block and turn it right grasp the purple block and turn it left
RT-1		
	VERB → NOUN → NOUN → ADP → ADJ → NOUN	place rxbar blueberry into bottom drawer move rxbar chocolate near orange can move 7up can near green can
	VERB → NOUN → NOUN → ADP → NOUN → NOUN	move water bottle near rxbar chocolate move coke can near water bottle move rxbar blueberry near water bottle
	VERB → NOUN → NOUN → ADP → ADJ → NOUN → CCONJ → VERB → ADP → NOUN	pick coke can from bottom drawer and place on counter pick water bottle from top drawer and place on counter pick rxbar blueberry from middle drawer and place on counter

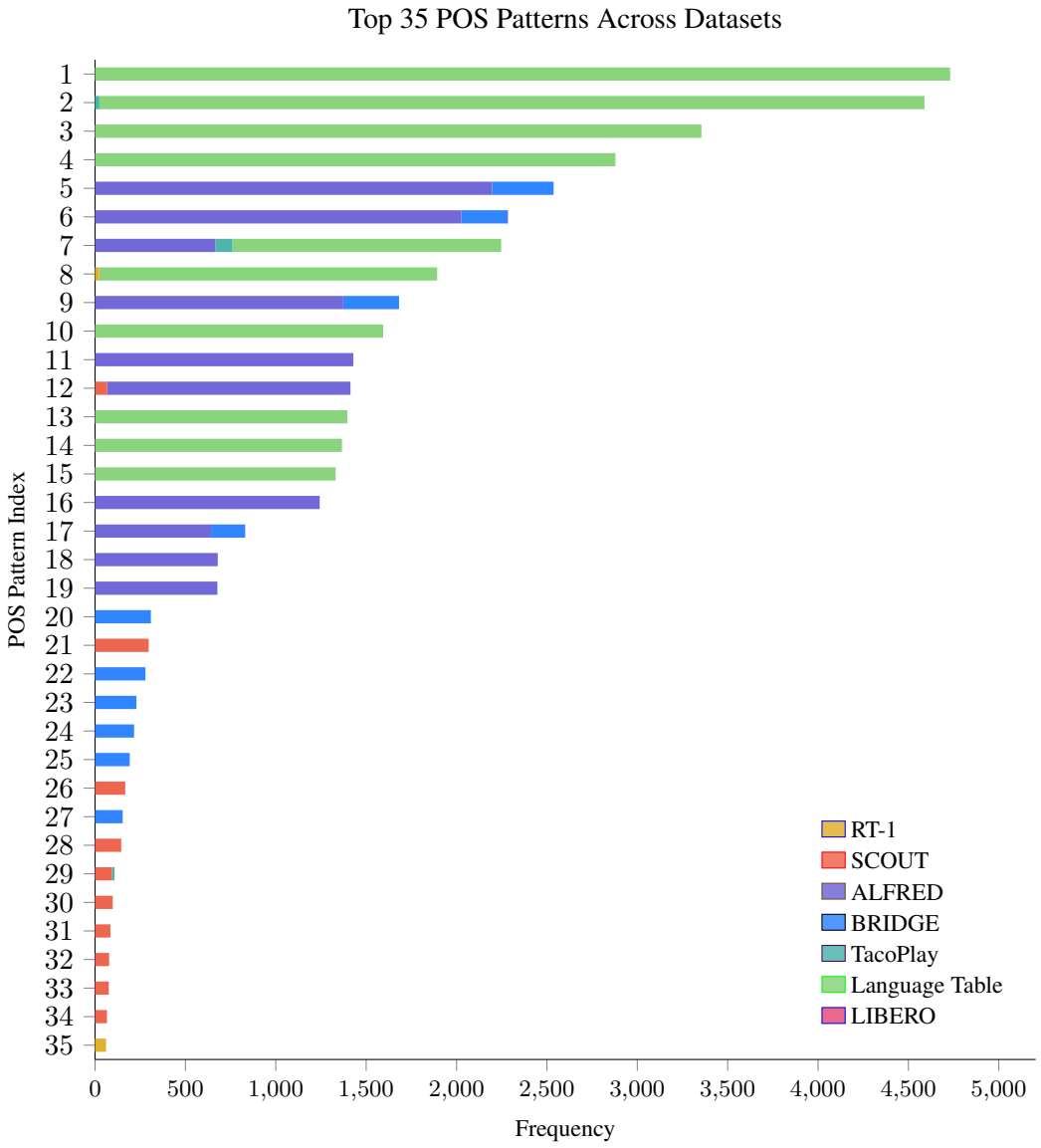
Figure 11: **Analysis 3: Structural Diversity** POS patterns and example sentences from TacoPlay and RT-1. Each example sentence is aligned with its corresponding POS pattern, and grouped by dataset.

Dataset	IAA H-H	Negation		Conditional		Multi-Step		Cycle	
		SE	IAA H-L	SE	IAA H-L	SE	IAA H-L	SE	IAA H-L
ALFRED	0.66	0.00	1.00	0.00	1.00	0.04 ± 0.02	0.92	0.01 ± 0.01	0.87
SCOUT	0.88	0.00	1.00	0.00	1.00	0.17 ± 0.03	0.20	0.00 ± 0.01	0.95
BRIDGE	0.99	0.00 ± 0.01	0.50	0.00	1.00	0.10 ± 0.03	0.72	0.01 ± 0.01	0.66
LanguageTable	0.55	0.00	1.00	0.00	1.00	0.14 ± 0.03	0.29	0.01 ± 0.01	0.90

Table 9: For datasets with more than 600 unique instructions, we randomly sampled 500 unique commands for manual annotation and calculated the following metrics. Inter-annotator agreement (IAA) as Cohen's κ and Standard Errors (SE) are reported. Human-Human IAA is denoted by H-H. Human-LLM IAA is denoted by H-L.

Dataset	**POS Pattern**	**Example Sentences**
SCOUT		
	VERB → ADV → NUM → NOUN	turn left thirty degrees turn left ninety degrees move forward one foot
	VERB → ADP → DET → NOUN	move towards a shoe move towards the barrel go through the door
	VERB → NUM → NOUN → ADV	turn sixty degrees left move ten inches northeast move two feet forward
BRIDGE		
	VERB → DET → NOUN → ADP → DET → NOUN → PUNCT	Place the mushroom behind the spatula. Place the salmon in the pot. Move the mushroom onto the towel.
	VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT	Move the spatula at the edge of the table. Move the spoon to the left of the napkin. Put the cloth to the left of the spoon.
	VERB → DET → NOUN → ADP → DET → ADJ → NOUN → PUNCT	Place the strawberry in the silver pot. Set the pot onto the green cloth. Place the pot on the blue cloth.

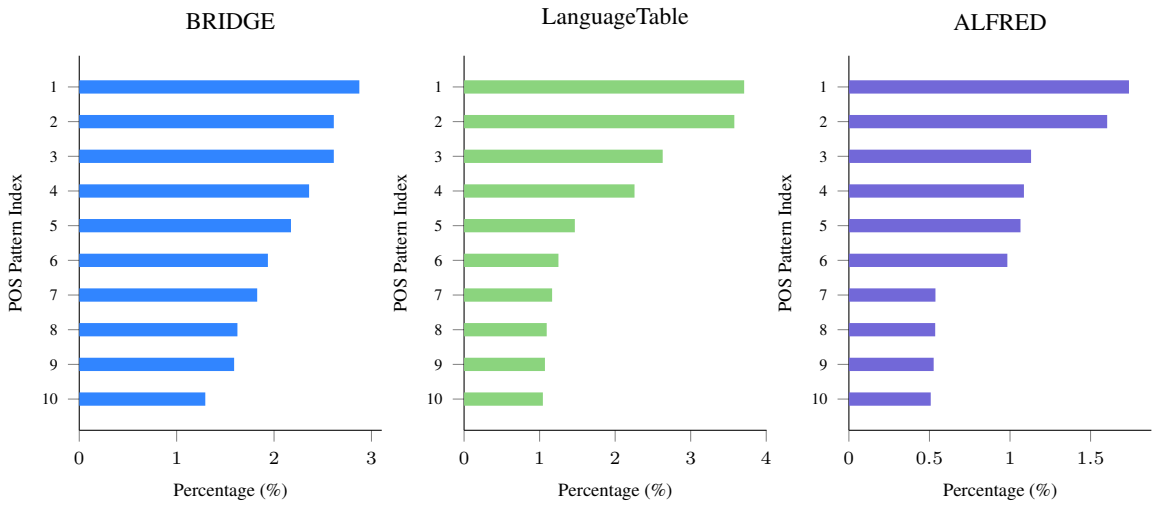
Figure 12: **Analysis 3: Structural Diversity** POS patterns and example sentences from SCOUT and BRIDGE datasets. Each example sentence is aligned with its corresponding POS pattern, and grouped by dataset.



POS Pattern Key:

1 VERB → DET → ADJ → NOUN → ADV → ADP → DET → ADJ → NOUN
2 VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN
3 VERB → DET → ADJ → NOUN → ADP → DET → NOUN → ADP → DET → ADJ → NOUN
4 VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN
5 VERB → DET → NOUN → ADP → DET → NOUN → PUNCT
6 VERB → DET → NOUN → ADP → DET → NOUN
7 VERB → DET → ADJ → NOUN → ADP → DET → NOUN
8 VERB → ADJ → NOUN → ADP → ADJ → NOUN
9 VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
10 VERB → DET → ADJ → NOUN → ADP → DET → NOUN → NOUN → ADP → DET → ADJ → NOUN
11 VERB → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
12 VERB → ADP → DET → NOUN → ADP → DET → NOUN
13 VERB → DET → ADJ → NOUN → ADP → ADJ → NOUN
14 VERB → DET → ADJ → NOUN → ADP → DET → ADJ → ADJ → NOUN → ADP → DET → ADJ → NOUN
15 VERB → DET → ADJ → NOUN → ADJ → ADP → DET → ADJ → NOUN
16 VERB → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
17 VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN
18 VERB → ADP → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
19 VERB → ADP → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
20 VERB → DET → NOUN → ADP → DET → ADJ → NOUN → PUNCT
21 VERB → ADV → NUM → NOUN
22 VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN → PUNCT
23 VERB → DET → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
24 VERB → DET → NOUN → ADP → DET → ADJ → NOUN
25 VERB → DET → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → NOUN
26 VERB → ADP → DET → NOUN
27 VERB → DET → NOUN → ADP → DET → ADJ → ADJ → NOUN → ADP → DET → NOUN
28 VERB → NUM → NOUN → ADV
29 VERB → ADP → DET → ADJ → NOUN
30 VERB → ADV
31 VERB → NOUN
32 NUM → NOUN
33 VERB → NUM → NOUN
34 VERB → ADP → NOUN
35 VERB → NOUN → NOUN → ADP → ADJ → NOUN

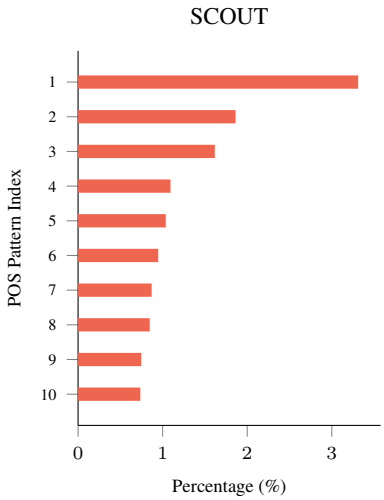
Figure 13: Top 35 POS patterns across datasets after aggregating the top 10 POS patterns within each dataset.



	BRIDGE
1	VERB → DET → NOUN → ADP → DET → NOUN → PUNCT
2	VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
3	VERB → DET → NOUN → ADP → DET → ADJ → NOUN → PUNCT
4	VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN → PUNCT
5	VERB → DET → NOUN → ADP → DET → NOUN
6	VERB → DET → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
7	VERB → DET → NOUN → ADP → DET → ADJ → NOUN
8	VERB → DET → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → NOUN
9	VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN
10	VERB → DET → NOUN → ADP → DET → ADJ → ADJ → NOUN → ADP → DET → NOUN

	LanguageTable
1	VERB → DET → ADJ → NOUN → ADV → ADP → DET → ADJ → NOUN
2	VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN
3	VERB → DET → ADJ → NOUN → ADP → DET → NOUN → ADP → DET → ADJ → NOUN
4	VERB → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN → ADP → DET → ADJ → NOUN
5	VERB → ADJ → NOUN → ADP → ADJ → NOUN
6	VERB → DET → ADJ → NOUN → ADP → DET → NOUN → NOUN → ADP → DET → ADJ → NOUN
7	VERB → DET → ADJ → NOUN → ADP → DET → NOUN
8	VERB → DET → ADJ → NOUN → ADP → ADJ → NOUN
9	VERB → DET → ADJ → NOUN → ADP → DET → ADJ → ADJ → NOUN → ADP → DET → ADJ → NOUN
10	VERB → DET → ADJ → NOUN → ADJ → ADP → DET → ADJ → NOUN

	ALFRED
1	VERB → DET → NOUN → ADP → DET → NOUN → PUNCT
2	VERB → DET → NOUN → ADP → DET → NOUN
3	VERB → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
4	VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
5	VERB → ADP → DET → NOUN → ADP → DET → NOUN
6	VERB → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
7	VERB → ADP → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN → PUNCT
8	VERB → ADP → DET → ADJ → NOUN → ADP → DET → NOUN → PUNCT
9	VERB → DET → ADJ → NOUN → ADP → DET → NOUN
10	VERB → DET → NOUN → ADP → DET → NOUN → ADP → DET → NOUN



	SCOUT
1	VERB → ADV → NUM → NOUN
2	VERB → ADP → DET → NOUN
3	VERB → NUM → NOUN → ADV
4	VERB → ADV
5	VERB → ADP → DET → ADJ → NOUN
6	VERB → NOUN
7	NUM → NOUN
8	VERB → NUM → NOUN
9	VERB → ADP → DET → NOUN → ADP → DET → NOUN
10	VERB → ADP → NOUN

Figure 14: Top 10 POS patterns by dataset (BRIDGE, LanguageTable, ALFRED, and SCOUT), normalized by dataset size.

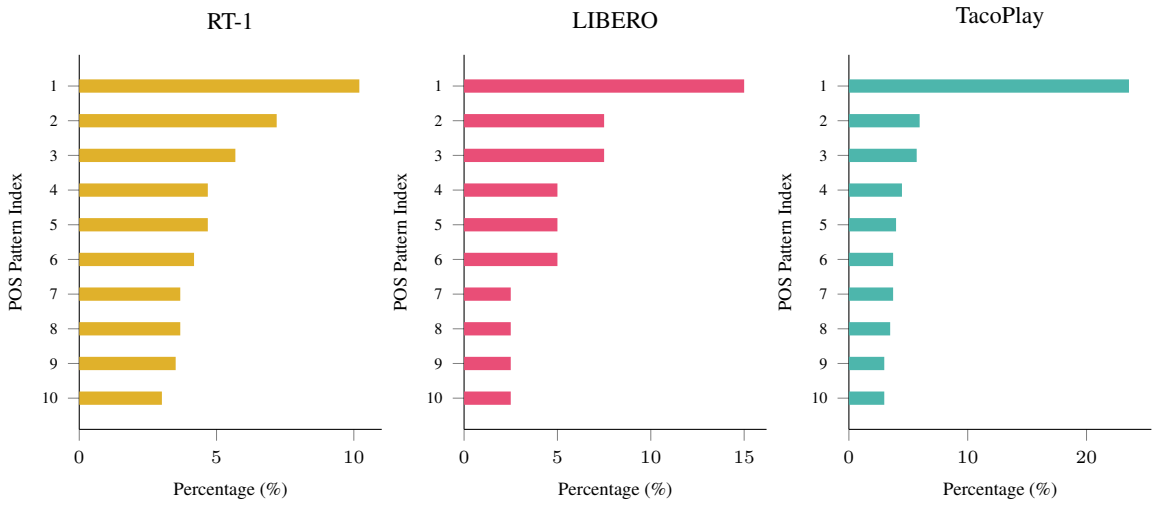


Figure 15: Top 10 POS patterns by dataset (RT-1, LIBERO, TACOPLAY), normalized by dataset size.

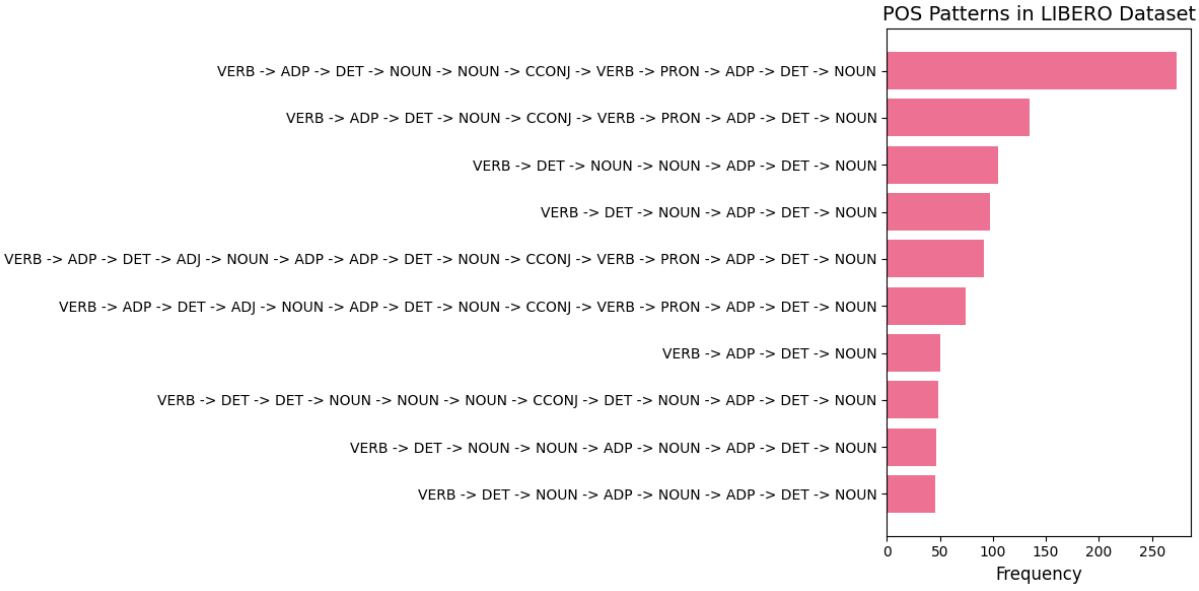


Figure 16: **Analysis 3: Structural Diversity** Dependency parse features across **all** LIBERO splits.

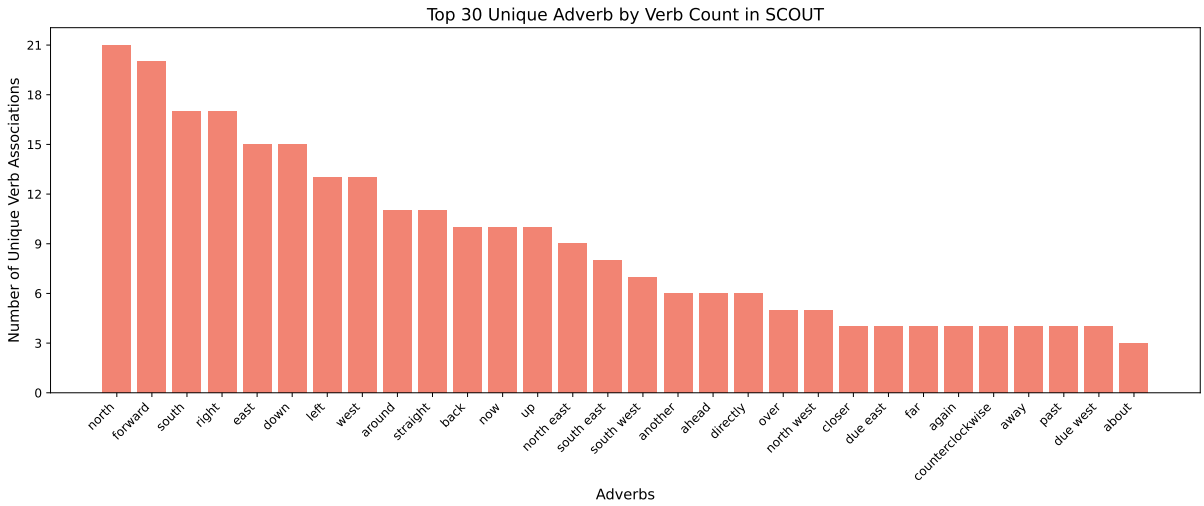


Figure 17: **Analysis 2: Semantic Diversity** VLN adverbials - limited to the top 30 adverbs with most unique language use

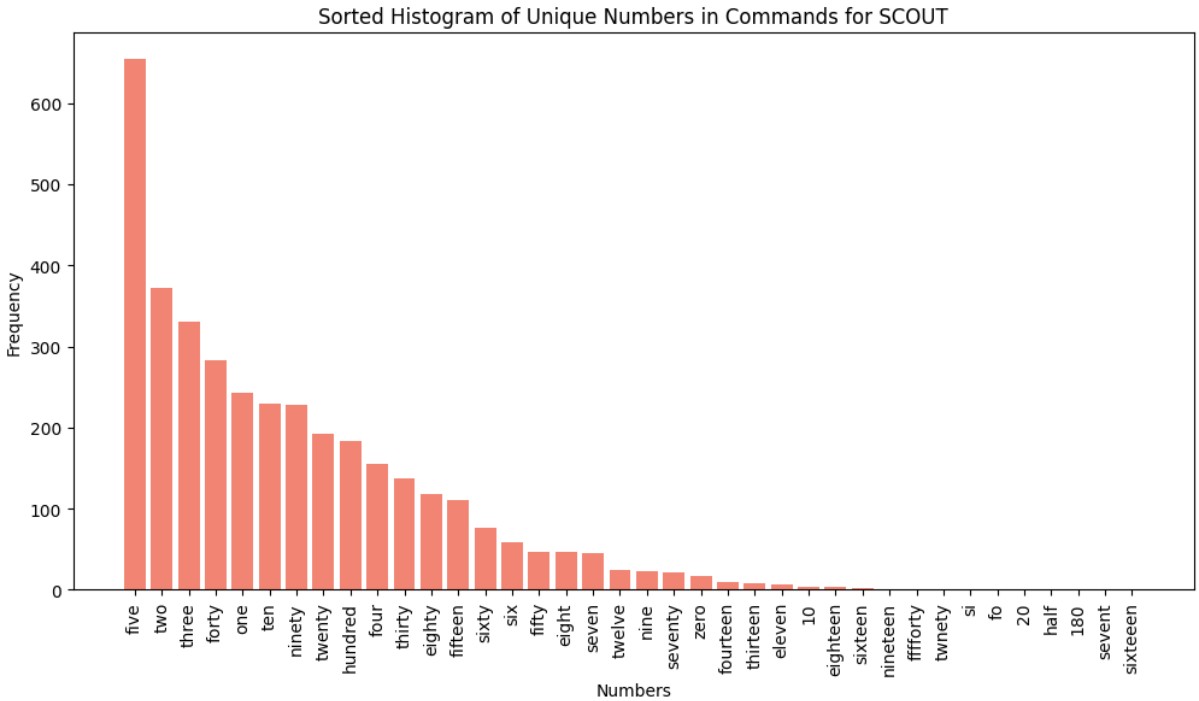


Figure 18: **Analysis 2: Semantic Diversity** SCOUT Numerics

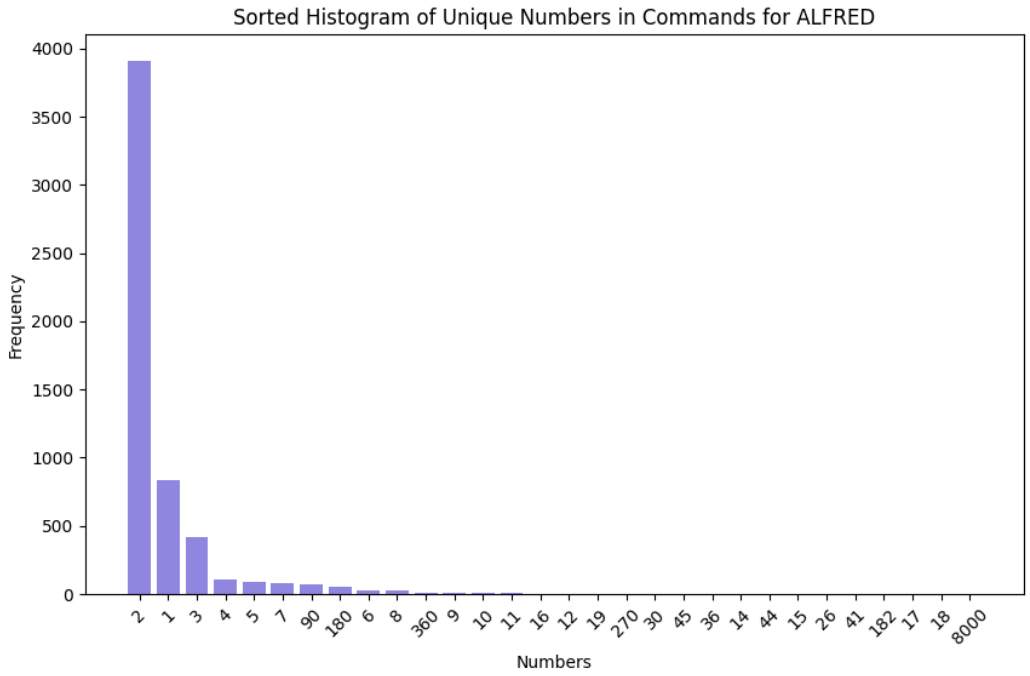


Figure 19: **Analysis 2: Semantic Diversity** ALFRED Numerics

Figure 20: Numeric representation in navigation datasets.

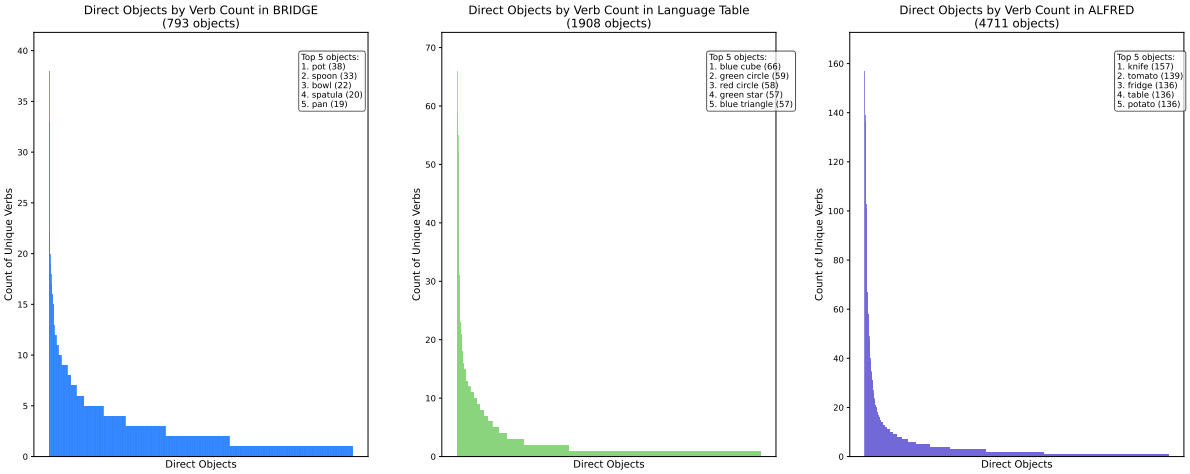


Figure 21: **Analysis 2: Semantic Diversity** Frequency Plot of Unique Verbs per Direct Object for Manipulation Datasets

Figure 22: **Analysis 2: Semantic Diversity** Frequency Plot of Unique Verbs per Direct Object for Manipulation Datasets

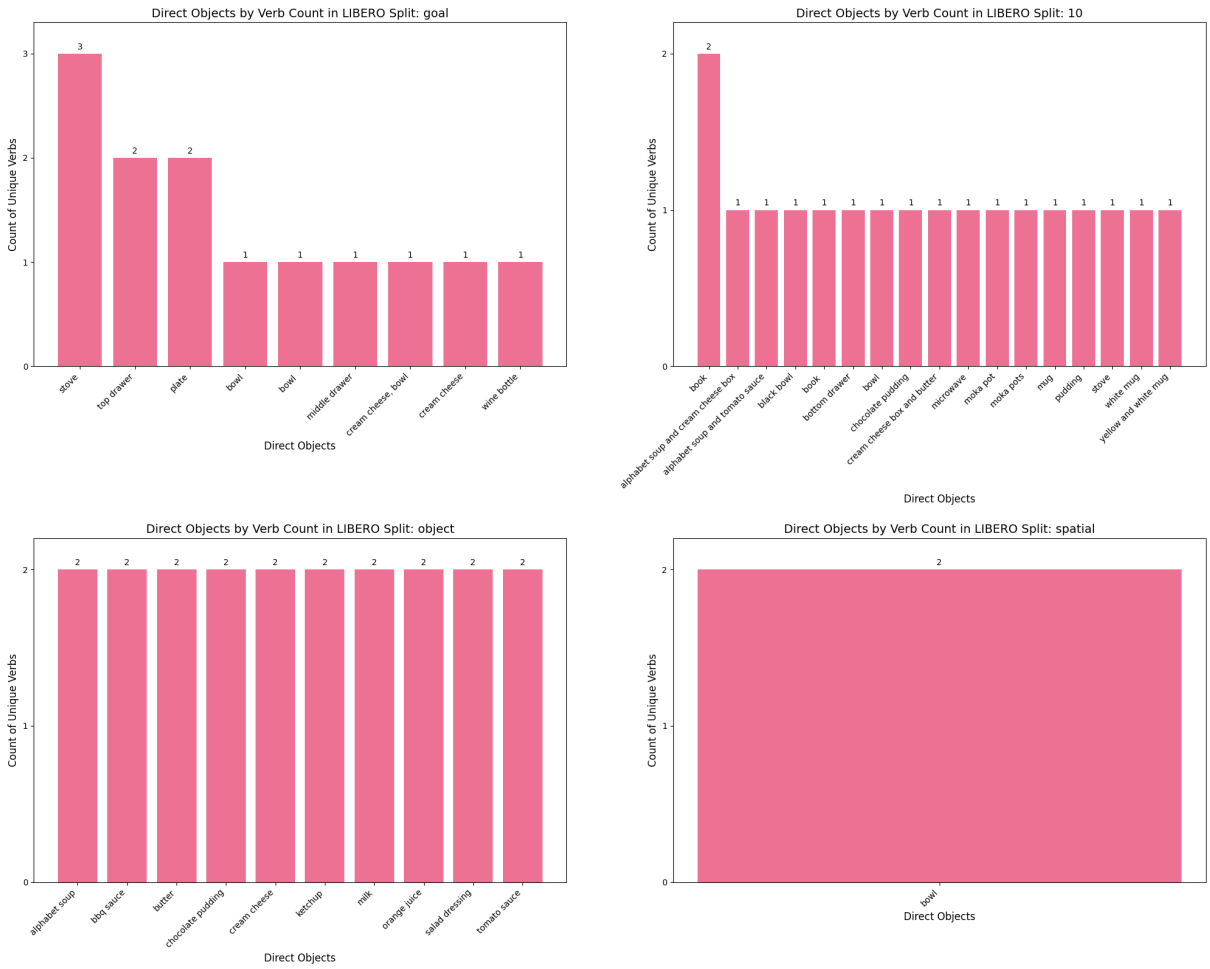


Figure 23: **Analysis 2: Semantic Diversity** Verb and direct object frequencies across **all** LIBERO splits.

Category	Dataset	Examples
Negation	SCOUT	i don't know what the red thing was you are not at the total entrance no i did not see any
	BRIDGE	video frames not showing video frames or not showing Picture is not downloading, not able to view.
	ALFRED	This step does not exist. Slice the tomato on the counter but do not put down the knife. Cook the potato slice in the microwave and do not put the cooked potato slice on the counter.
Conditional	SCOUT	see if there's a doorway check and see if there's a doorway there and i'll point out when there's a doorway so we can count them
	BRIDGE	Pick the orange towel and place it on the middle if the table PLACE THE YELLOW TOPWEL SIDE IF THE TABLE
	ALFRED	Take keys from the black table, leave them on the lamp when you turn it on. Turn right and walk until you're even with the fridge on your right and when you are turn right and walk to it. Turn left and walk to the table then turn right when you get to it.
Multi-Step	LIBERO	open the top drawer and put the bowl inside
	TacoPlay	go towards the drawer and place the pink object go towards the purple block and grasp it take the purple block and rotate it right
	RT-1	pick coke can from bottom drawer and place on counter pick apple from top drawer and place on counter pick green rice chip bag from bottom drawer and place on counter
	SCOUT	and take a picture and then the last question here anything that indicates the environment was recently occupied and then take a picture
	BRIDGE	put pot or pan on stove and put egg in pot or pan Take the spatula from the vessel and place it on the table.
	ALFRED	Open the drawer. Put the cell phone in the drawer on the right side towards the back and close it. open the top right drawer of the desk, put phone inside, close the drawer Turn and move to the far end of the kitchen island, so you're facing the tomato and fork.
Cycle	SCOUT	continue moving forward follow hallway to the end of the wall uh to until you reach the wall take a photo every forty five degrees
	BRIDGE	end effector reaching knife pick orange toy from vessel and keep it on the left side of the table end effector reaching corn
	ALFRED	Move over to the right side of the desk again. Put the potato slice in the fridge and shut the door and then take the potato slice out and shut the fridge door again. Walk to your left until you see a loaf of bread on the counter top.

Table 10: **Analysis 3: Structural Diversity** Representative instruction examples for negation, conditional, multi-step, and cycle structures. Note that in BRIDGE and ALFRED, some examples contain noise from the original OXE metadata (e.g., typos or syntactic errors); and in many cases, this noise artificially inflate diversity scores.