Scaling Diverse Language Generation for 3D Visual Grounding

Austin T. Wang[1] Dongchen Yang[1] Angel X. Chang[1,2]
[1]Simon Fraser University [2]Amii

Defining Diversity

We group 3DVG prompts by their *targets/anchors*, *attributes*, *relationships*, and other language structures.

Find the food storage which does not have a green, rectangular object on top of it.

Imperative | Target | Negation | Color | Shape | Anchor | Vert. | Coref.

Target Reference
- Type (generic, coarse-grained, fine-grained)
- Coreferences
- Noun phrase position

Attribute Understanding
- Counts (total, target, anchor)
- Types

Relationship Understanding
- Counts (total, target, anchor)
- Types

Anchor Reference
- Type (single, multiple, non-object, agent-based)

Diversity Statistics
- Lexical bigrams

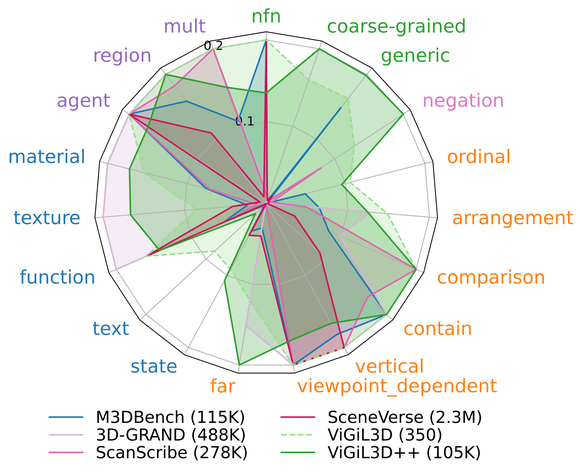
Structural Patterns
- Sentence structure (e.g. imperative, inverted)
- Logical operators (e.g. negation)

Attribute Types	
Color	Function
Size	Texture
Shape	Style
Number	Text Label
Material	State

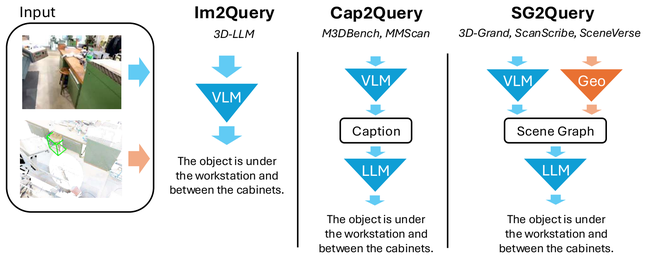
Relationship Types	
Near	Contain
Far	Arrangement
Directional	Ordinal
Vertical	Comparison

How diverse are existing datasets?

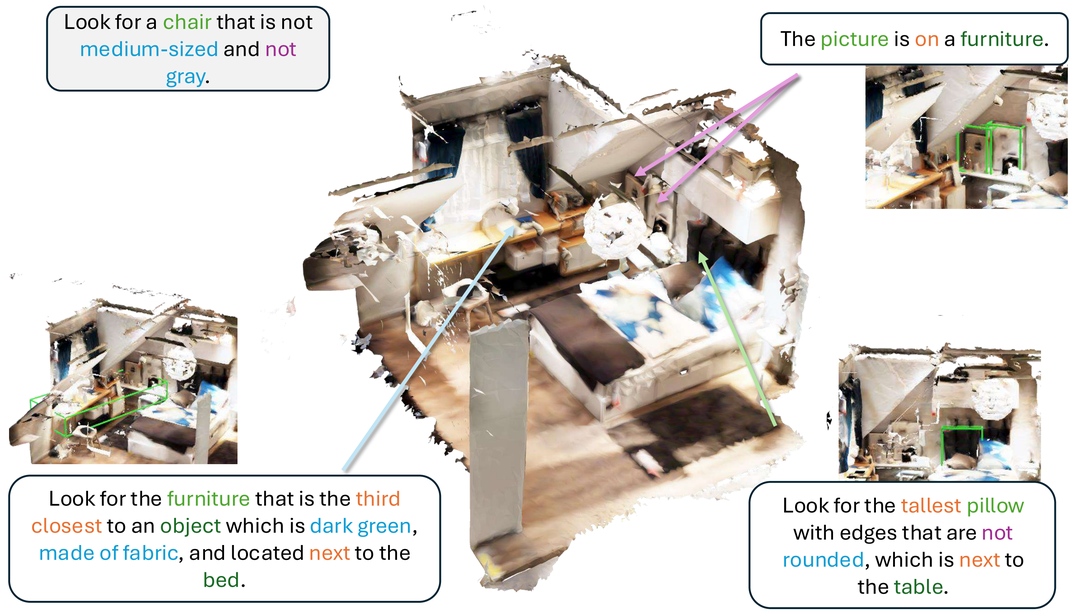
ViGiL3D++ has equal or greater linguistic diversity than existing scaled datasets.



Baselines



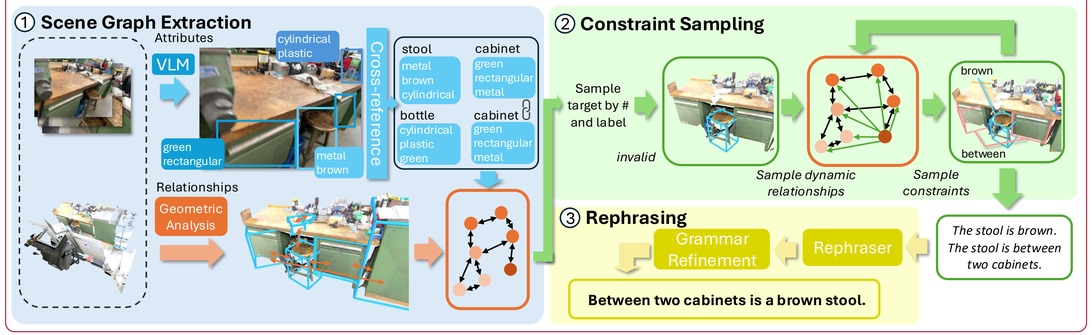
How important is the diversity of the 3DVG descriptions we use for training and evaluation?



Can VLMs be used for 3DVG query generation?

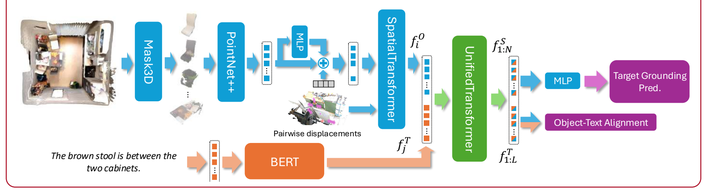
ViGiL3D++:

1) Use *cross-referencing* to extract a more consistent scene graph
2) Use *constraint sampling* for diversity and logical consistency in queries



V3DM Model

We train a model based on 3D-VisTA [9] to test the utility of ViGiL3D++.



Training on Diverse Data

Training on ViGiL3D++ yields better performance on 3DVG benchmarks.

Model	Data: Pretrain	Data: Finetune	ViGiL3D [6]: Acc	ViGiL3D [6]: F1	Acc on ScanRefer [1]: Uniq	Acc on ScanRefer [1]: Multi	Acc on ScanRefer [1]: Overall
ZSVG3D [8]	–	–	18.9	12.2	73.9	24.9	34.4
3D-VisTA [9]	ScanScribe	ScanRefer	15.1	15.1	84.6	**48.2**	55.2
3D-GRAND [7]	3D-GRAND	ScanRefer	16.0	9.2	73.3	33.0	40.8
GPS [3]	SceneVerse	ScanRefer	**24.1**	14.7	88.1	46.4	54.5
V3DM	ScanScribe	ViGiL3D++$_{SR}$	22.9	**15.7**	**88.7**	47.6	**55.6**

Insights

- ◇ Training and evaluating on diverse language descriptions is important.
- ◇ VLMs do not reliably reason about spatial constraints (preserving consistency and generating with diversity).
- ◇ Encoding intuition about spatial relationships into concrete scene graphs is challenging due to the contextual information required.

Acknowledgements: This work was funded in part by a CIFAR AI Chair and the NSERC Discovery Grant, and enabled by support from the Digital Research Alliance of Canada and a CFI/BCKDF JELF.

References:

[1] D. Z. Chen et al. ScanRefer. In *ECCV*, 2020.
[2] Y. Hong et al. 3D-LLM. *NeurIPS*, 2023.
[3] B. Jia et al. SceneVerse. In *ECCV*, 2024.
[4] M. Li et al. M3DBench. In *ECCV*, 2024.
[5] R. Lyu et al. MMScan. *NeurIPS*, 2024.
[6] A. T. Wang et al. ViGiL3D. In *ACL*, 2025.
[7] J. Yang et al. 3D-GRAND. In *CVPR*, 2025.
[8] Z. Yuan et al. ZSVG3D. In *CVPR*, 2024.
[9] Z. Zhu et al. 3D-VisTA. In *ICCV*, 2023.