Voxel-informed Language Grounding

Anonymous ACL submission

Abstract

Even when applied to 2D images, natural language describes a fundamentally 3D world. We present the Voxel-informed Language Grounder (VLG), a language grounding model that leverages *3D geometric information* in the form of voxel maps derived from the visual input using a volumetric reconstruction model. We show that VLG significantly improves grounding accuracy on SNARE (Thomason et al., 2021), an object reference game task. At the time of writing, VLG holds the top place (anonymized) on the SNARE leaderboard[1], achieving SOTA results with a 1.7% overall improvement on all descriptions.

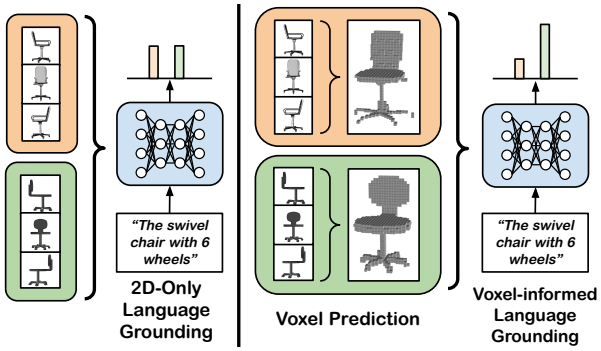

Figure 1: **Voxel-informed Language Grounder.** Our VLG model leverages explicit 3D information by inferring volumetric voxel maps from input images, allowing the agent to reason jointly over the geometric and visual properties of objects when grounding.

1 Introduction

Embodied robotic agents hold great potential for providing assistive technologies in home environments (Pineau et al., 2003), and natural language provides an intuitive interface for users to interact with such systems (Andreas et al., 2020). For these systems to be effective, they must be able to reliably ground language in perception (Bisk et al., 2020; Bender and Koller, 2020).

Despite typically being paired with 2D images, natural language that is grounded in vision describes a fundamentally 3D world. For example, consider the grounding task in Figure 1, where the agent must select a target chair against a distractor given the description "the swivel chair with 6 wheels." Although the agent is provided with multiple images revealing all of the wheels on each chair, it must be able to properly aggregate information across images to successfully differentiate them, something that requires reasoning about their *3D geometry* at some level.

In this work we show how language grounding performance may be improved by leveraging 3D prior knowledge. Our model, Voxel-informed Language Grounder (VLG), extracts 3D voxel maps using a pre-trained *volumetric reconstruction model*, which it fuses with multimodal features from a large-scale vision and language model in order to reason jointly over the visual and 3D geometric properties of objects.

We focus our investigation within the context of SNARE (Thomason et al., 2021), an object reference game where an agent must ground natural language describing common household objects by their *geometric* and visual properties, showing that grounding accuracy significantly improves by incorporating information from predicted 3D volumes of objects. At the time of writing, VLG achieves SOTA performance on SNARE, attaining an absolute improvement of 1.7% over the next closest baseline.

2 Related Work

Prior work has studied deriving structured representations from images to scaffold language grounding. However, a majority of systems use representations such as 2D regions of interest (Anderson et al., 2018; Wang et al., 2020) or symbolic graph-based representations (Hudson and Manning, 2019; Kulkarni et al., 2013), which do not encode 3D

[1] https://github.com/snaredataset/snareleaderboard

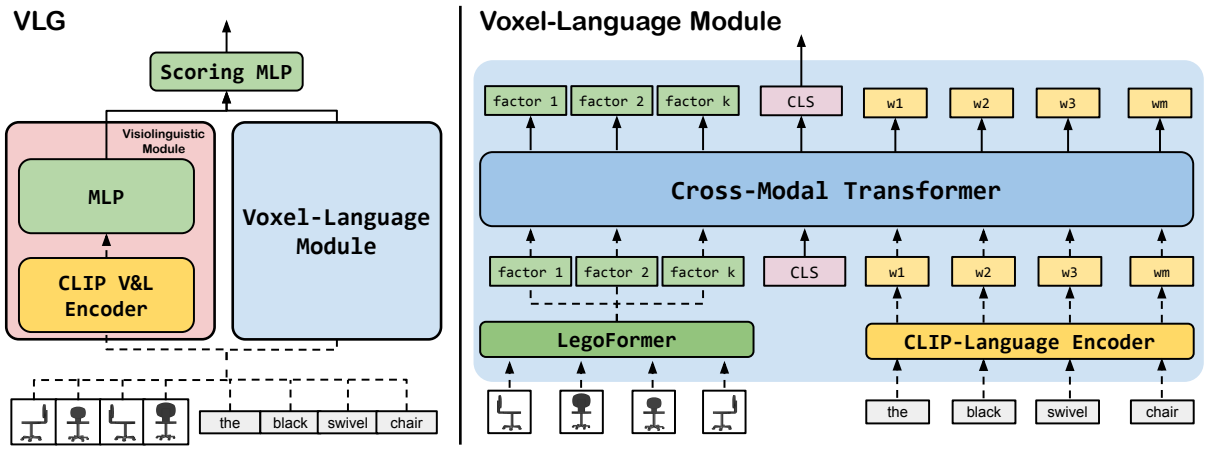

Figure 2: **VLG Architecture.** (Left) Our VLG model consists of a visiolinguistic module which produces a joint embedding for text and images using CLIP and a voxel-language module for jointly embedding language and volumetric maps. (Right) The voxel-language module uses a cross modal transformer to fuse word embeddings from CLIP with voxelmap factors extracted from LegoFormer (Yagubbayli et al., 2021). During training, gradients only flow through solid lines.

properties of objects.

Most prior work tying language to 3D representations has largely focused on generating 3D structures conditioned on language, either at the scene (Chang et al., 2014, 2015a), pose (Ahuja and Morency, 2019; Lin et al., 2018), or object (Chen et al., 2018) level. In contrast, in this work we focus on augmenting language grounding using structured 3D representations derived from 2D images. For the task of visual language navigation, prior work has shown how a persistent 3D semantic map may be used as an intermediate representation to aid in selecting navigational waypoints (Chaplot et al., 2020; Blukis et al., 2021). The semantic maps, however, represent entire scenes with voxels representing object categories, rather than their geometric properties. In this work, we show how a more granular occupancy map representing objects' geometry can improve language grounding.

Closest to our work is that of Prabhudesai et al. (2020), which presents a method for mapping language to 3D features within scenes from the CLEVR (Johnson et al., 2017) dataset. Their system generates 3D feature maps inferred from images and then grounds language directly to 3D bounding boxes or coordinates. Their system assumes, however, that dependency trees are provided for the natural language, and it is trained with supervised alignments between tree constituents and the 3D representations.

3 Voxel-informed Language Grounder

We consider a task where an agent must correctly predict a target object v^t against a distractor v^c given a natural language description $w^t = \{w_1, ..., w_m\}$ of the target. For each object, the agent is provided with n 2D views $v = \{x_1, ..., x_n\}$, $x_i \in \mathcal{R}^{3\times W\times H}$.

An agent for this task is represented by a scoring function $s(v, w) \in [0, 1]$, computing the compatibility between the target description and the 2D views of an object. We first use unimodal encoders to encode the language description into $e_w = h(w)$ and the object view images into a single aggregate visual embedding $e_v = g(v)$ before fusing them with a visiolinguistic module $e_{vw} = f_{vw}([e_v; e_w])$. Prior approaches to this problem directly input this fused representation to a scoring module to produce a score $s(e_{vw})$. They do not explicitly reason about the 3D properties of the observed objects, requiring the models to learn them implicitly.

In contrast, our Voxel-informed Language Grounder augments the scoring function s with explicit 3D volumetric information $e_o = o(v)$ extracted from a pre-trained multiview reconstruction model $o(v)$. The volumetric information (in the form of a voxel occupancy map in $\mathcal{R}^{W\times H\times D}$) is first fused into a joint representation with the language using a multimodal voxel-language module $e_{ow} = f_{ow}([e_o; e_w])$. The scoring function then produces a score based on all three modalities $s([e_{vw}; e_{ow}])$.

	VALIDATION			TEST		
Model	Visual	Blind	All	Visual	Blind	All
ViLBERT	89.5	76.6	83.1	80.2	**73**	76.6
MATCH	89.2 (0.9)	75.2 (0.7)	82.2 (0.4)	83.9 (0.5)	68.7 (0.9)	76.5 (0.5)
MATCH*	90.6 (0.004)	75.7 (0.01)	83.2 (0.006)	-	-	-
LAGOR	89.8 (0.4)	75.3 (0.7)	82.6 (0.4)	84.3 (0.4)	69.4 (0.5)	77.0 (0.5)
LAGOR*	89.6 (0.003)	74.9 (0.003)	82.3 (0.0)	-	-	-
VLG (Ours)	**91.6** (0.008)	**78.5**†(0.002)	**85.2**†(0.004)	**85.8**	71.3	**78.7**

Table 1: **SNARE Benchmark Performance.** Object reference game accuracy on the SNARE task across validation and test sets. Performance on models with an asterisk are our replications of the baselines in (Thomason et al., 2021). MATCH*, LAGOR*, and VLG performances are averaged over 3 seeds. Standard deviations are shown in parentheses. Our VLG model achieves the best overall performance. Due to leaderboard submission restrictions, we were not able to get test set results for the MATCH* and LAGOR* replications. † denotes statistical significance in improvement over the next best model (with $p < 0.05$).

3.1 Model Architecture

Visiolinguistic Module. The architecture of our visiolinguistic module f_{vw} (on left panel, Figure 2) largely mirrors the architecture of MATCH from (Thomason et al., 2021). A pre-trained CLIP-ViT (Radford et al., 2021) model is used to encode the language description and view images into vectors in $\mathcal{R}^{512}$. The image embeddings are max-pooled and concatenated to the description embedding before being passed into an MLP which generates a fused representation.

Voxel-Language Module. We use representations extracted from a ShapeNet (Chang et al., 2015b; Wu et al., 2015) pre-trained LegoFormerM (Yagubbayli et al., 2021), a multi-view 3D volumetric reconstruction model, as input to our voxel-language module f_{ow}. LegoFormer[2] is a transformer (Vaswani et al., 2017) based model whose decoder generates volumetric maps factorized into 12 parts. Each object factor is represented by a set of three vectors $x, y, z \in \mathcal{R}^{32}$, which we concatenate to use as input tokens for our voxel-language module. A triple cross-product over x, y, z may be used to recover a 3D volume $\mathcal{V} \in \mathcal{R}^{32\times32\times32}$ for each factor. The full volume for the object is generated by aggregating the factor volumes through a sum operation. For more details on LegoFormer, we refer the reader to (Yagubbayli et al., 2021).

We use a cross-modal transformer (Vaswani et al., 2017) encoder to fuse the language and object factors (Figure 2, right). The cross-modal transformer takes as input language tokens, in the form of CLIP word embeddings, and the 12 object factors output by the LegoFormer decoder, which contain the inferred geometric occupancy information of the object. We use a CLS token as an aggregate representation of the language and object factors.

The final scoring layer of our model is represented by an MLP which takes as input the concatenation of the visiolinguistic model output and the cross-modal transformer's CLS token.

4 Language Grounding Evaluation

We test our method on the SNARE benchmark (Thomason et al., 2021). SNARE is a language grounding dataset which augments ACRONYM (Eppner et al., 2021), a grasping dataset built off of ShapeNetSem (Savva et al., 2015; Chang et al., 2015a), with natural language annotations of objects.

SNARE presents an object reference game where an agent must correctly guess a target object against a distractor. In each instance of the game, the agent is provided with a language description of the target as well as multiple 2D views of each object. SNARE differentiates between **visual** and **blind** object descriptions. For visual descriptions, AMT workers were primed to describe objects by *name*, *shape*, and *color* (e.g. "classic armchair with white seat"). In contrast, for blind descriptions workers were primed to describe objects by *shape* and *parts* (e.g. "oval back and vertical legs") in order to get descriptions biased towards objects' geometric properties. The train/validation/test sets were generated by splitting over (207 / 7 / 48) ShapeNetSem object categories, respectively containing (6,153 / 371 / 1,357) unique object instances and (39,104 / 2,304 / 8,751) object pairings with referring expressions. Renderings are provided for each object

[2] https://github.com/faridyagubbayli/LegoFormer

Model	Visual	Blind	All
VGG16	91.6 (0.004)	75.9 (0.008)	83.8 (0.006)
MLP	91.2 (0.007)	77.7 (0.007)	84.5 (0.007)
no-CLIP	67.7 (0.006)	69.0 (0.007)	68.4 (0.002)
VLG	91.6 (0.008)	**78.5** (0.002)	**85.2** (0.004)

Table 2: **Ablation Study.** SNARE reference game accuracy across ablations of our model on the validation set. Performance is averaged over 3 seeds for each condition, with standard deviations in parentheticals.

instance over 8 canonical viewing angles.

We compare VLG against the set of models provided with SNARE. At the time of writing, these were the only available models for the task. All SNARE[3] baselines except **ViLBERT** use a CLIP-ViT (Radford et al., 2021) backbone for encoding both images and language descriptions. We refer the reader to Appendix A.1 for details.

5 Results

We present average performance for trained models over 3 seeds with standard deviations on the validation set. We also present test set performance for VLG and the performance of the SNARE baselines reported by Thomason et al. (2021) (See Appendix A.2 for details on training procedures).

5.1 Comparison to SOTA

In Table 1 we can observe reference game performance for all models. VLG achieves SOTA performance with an absolute improvement on the test set of 1.7% over LAGOR, the next best leaderboard model. Although there is a general improvement of 1.5% in **visual** reference grounding, there is an improvement of 1.9% in **blind** reference grounding. This suggests that the injected 3D information is more useful for disambiguating between examples referring to geometric properties of the referred objects. Improvements on the Blind and All conditions of the validation set are statistically significant (with $p < 0.05$) under a Welch's two-tailed t-test.

5.2 Ablation Study

We present a variety of ablations on the validation set to investigate the contributions of each piece of our model. All results can be observed in Table 2.

VGG16 Embeddings. LegoFormer uses an ImageNet (Deng et al., 2009) pre-trained VGG16 (Simonyan and Zisserman, 2014) as a backbone for extracting visual representations, which is a different dataset and pre-training task than what the CLIP-ViT image encoder is trained on. This presents a confounding factor which we ablate by performing an experiment where we feed our model's scoring function VGG16 features directly instead of LegoFormer object factors (VGG16 in Table 2). Despite getting comparable results to VGG16 on visual reference grounding, VLG provides a clear improvement in blind (and therefore overall) reference performance, suggesting that the extracted 3D information is useful for grounding more geometrically based language descriptions, with the VGG16 features being largely redundant in terms of visual signal.

Architecture. We ablate the contribution of our cross-modal transformer branch by comparing it against an MLP mirroring the structure of the SNARE MATCH baseline. This model (MLP in Table 2) max-pools the LegoFormer object factors and concatenates the result to the CLIP visual and language features before passing them to an MLP scoring function. The MLP model overall outperforms the SNARE baselines from Table 1, corroborating the usefulness of the 3D information for grounding, but does not result in as large an improvement as the cross-modal transformer. This suggests that the transformer is better able at integrating information from the multi-view input.

CLIP Visual Embeddings. Finally, we evaluate the contribution of the visiolinguistic branch of the model by removing it and only using the cross-modal transformer over language and object factors. As may be observed, there is a large drop in performance, particularly for visual references. These results suggest that maintaining visual information such as color and texture is critical for good performance on this task, since the LegoFormer outputs contain only volumetric occupancy information.

6 Discussion

We have presented the Voxel-informed Language Grounder, a model which leverages explicit 3D information from predicted volumetric voxel maps to improve language grounding performance. VLG achieves SOTA results on SNARE, and ablations corroborate the effectiveness of using this 3D information for grounding. We hope this paper may encourage further work on integrating structured 3D representations into language grounding tasks.

[3] https://github.com/snaredataset/snare

References

Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. 2019. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8938–8947.

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. 2021. A persistent spatial semantic representation for high-level natural language instruction execution. *arXiv preprint arXiv:2107.05612*.

Angel Chang, Manolis Savva, and Christopher D Manning. 2014. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015a. Shapenet: An information-rich 3d model repository. Cite arxiv:1512.03012.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015b. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*, pages 100–116. Springer.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Clemens Eppner, Arsalan Mousavian, and Dieter Fox. 2021. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE.

Drew A Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903.

Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. 2018. Generating animated videos of human activities from natural language descriptions. *Learning*, 2018:1.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Joelle Pineau, Michael Montemerlo, Martha Pollack, Nicholas Roy, and Sebastian Thrun. 2003. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and autonomous systems*, 42(3-4):271–281.

Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W Harley, and Katerina Fragkiadaki. 2020. Embodied language grounding with 3d visual feature representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2220–2229.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Manolis Savva, Angel X. Chang, and Pat Hanrahan. 2015. Semantically-Enriched 3D Models for Common-sense Knowledge. *CVPR 2015 Workshop on Functionality, Physics, Intentionality and Causality*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. 2021. Language grounding with 3d objects. In *5th Annual Conference on Robot Learning*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ruocheng Wang, Jiayuan Mao, Samuel J Gershman, and Jiajun Wu. 2020. Language-mediated, object-centric representation learning. *arXiv preprint arXiv:2012.15814*.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920.

Farid Yagubbayli, Alessio Tonioni, and Federico Tombari. 2021. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*.

A Appendix

A.1 SNARE Baselines

Here we briefly describe the baselines provided by SNARE. For more details, we refer the reader to (Thomason et al., 2021).

MATCH uses a learned MLP to produce a score over CLIP-ViT language and pooled image embeddings.

ViLBERT fine-tunes a 12-in1 (Lu et al., 2020) pre-trained ViLBERT(Lu et al., 2019). This baseline is additionally provided with ground-truth image bounding boxes during training.

LAGOR. LAGOR's (**La**nguage **G**rounding through **O**bject **R**otation) scoring function mirrors the architecture of the MATCH module. During training, LAGOR is augmented with an auxiliary view-prediction loss, which tasks the agent with predicting the canonical view angle for each image given its embedding. LAGOR uses a separate MLP to produce view-predictions.

A.2 Training Procedure

We train each model for 75 epochs, reporting performance of the best performing checkpoint on the validation set. For the SNARE MATCH* and LAGOR* baselines we use the hyperparameters reported by Thomason et al. (2021). For all variants of our VLG model we use the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 1e-3, linear learning rate warmup of 10K steps, and a smoothed binary cross-entropy loss (Achlioptas et al., 2019). We use a computing cluster with RTX 2080 GPUs to run our experiments. All code to replicate our results will be made publicly available.