Natural Language Grounded Multitask Navigation

Xin Wang[1]*, Vihan Jain[2]*, Eugene Ie[2], William Yang Wang[1], Zornitsa Kozareva[2], Sujith Ravi[2]
[1]University of California, Santa Barbara [2]Google Research
{xwang,william}@cs.ucsb.edu, {vihanjain,eugeneie,kozareva,sravi}@google.com

Abstract

Recent research efforts enable the study of natural language grounded navigation in photo-realistic environments, e.g., following natural language instructions or dialog. However, data scarcity is a critical issue in these tasks, as conducting human demonstrated language interactions in the simulator is still expensive and time-consuming and it is impractical to exhaustively collect samples for all variants of the navigation tasks. Therefore, we introduce a generalized multitask navigation model that can seamlessly be trained on language-grounded navigation tasks such as Vision-Language Navigation (VLN) and Navigation from Dialog History (NDH). Benefiting from richer natural language guidance, the multitask model can efficiently transfer knowledge across related tasks. Experiments show that it outperforms the single-task model by 7% (success rate) on VLN and 61% (goal progress) on NDH, establishing the new state of the art for NDH.

1 Introduction

Navigation in visual environments by following natural language guidance [10] is a fundamental capability of intelligent robots that simulate human behaviors, because humans can easily reason about the language guidance and navigate efficiently by interacting with the visual environments. Recent efforts [3, 6, 19] empower large-scale learning of natural language grounded navigation that is situated in photo-realistic simulation environments.

Nevertheless, data scarcity is still a critical issue in these tasks. Unlike vision-only navigation tasks [16, 23, 15, 14] where episodes can be exhaustively sampled in simulation, natural language grounded navigation is supported by human demonstrated interaction in natural language. It is time-consuming and impractical to fully collect all the samples for individual tasks. Some data augmentation techniques [9, 18] have been proposed to alleviate data scarcity. For example, a speaker model can be trained to generate instructions on newly sampled trajectories [9], but natural language generation itself is extremely challenging. Recent work [11] indicates that machine-generated instructions suffer from the quality issue and thus only a small fraction of them is useful.

In this paper, we aim at resolving the data sparsity with multitask learning, leveraging and transferring knowledge across tasks. We propose a generalized multitask model for natural language grounded navigation tasks such as Vision-Language Navigation (VLN) and Navigation from Dialog History (NDH), where all the learnable parameters are shared and jointly trained on both tasks. We also adopt an interleaved multitask data sampling strategy to prevent the shared model from being dominated by one task. Experiments on VLN and NDH show that the multitask navigation model can not only efficiently execute different language guidance in indoor environments but also outperform the single-task baseline models by a large margin on both tasks. Together with specialized reward shaping, our model achieves new state-of-the-art performance on NDH. To the best of our knowledge, we introduce the first natural language grounded multitask navigation model and validate its effectiveness on VLN and NDH tasks.

[*] Equal contribution. Work was done when Xin Wang was interning at Google.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

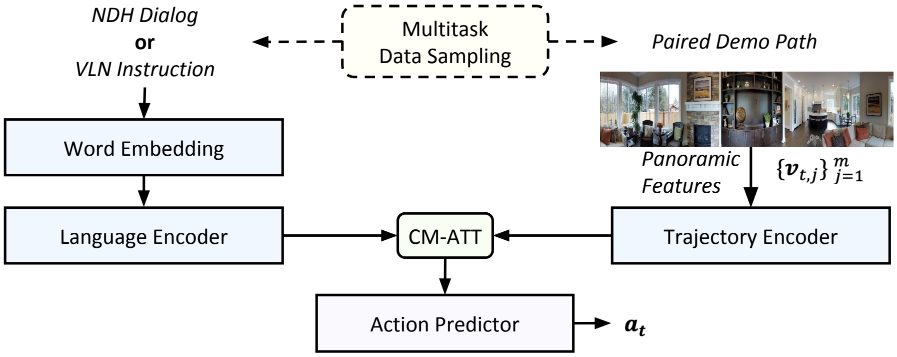

Figure 1: Natural Language Grounded Multitask Navigation Framework

2 Background

Vision-Language Navigation. Vision-Language Navigation [3, 5] task requires an embodied agent to navigate in photo-realistic environments to carry out natural language instructions. The agent is spawn at an initial pose $p_0 = (v_0, \phi_0, \theta_0)$, which includes the spatial location, heading and elevation angles. Given a natural language instruction $X = \{x_1, x_2, ..., x_n\}$, the agent is expected to perform a sequence of actions $\{a_1, a_2, ..., a_T\}$ and arrive at the target position v_{tar} specified by the language instruction X, which describes step-by-step instructions from the starting position to the target position. In this work, we consider VLN task defined for Room-to-Room (R2R) [3] dataset. The conventional metrics used to evaluate agent performance on VLN include success rate (SR), navigation error (NE), path length (PL), SPL [2] and CLS [13].

Navigation from Dialog History. Most recently, a Cooperative Vision-and-Dialog Navigation (CVDN) dataset [19] is introduced towards interactive language assistance for indoor navigation, which consists of over 2k embodied, human-human dialogs situated in photo-realistic home environments. The task of Navigation from Dialog History (NDH) is defined as: given a target object t_o and a dialog history between humans cooperating to perform the task, the embodied agent must infer navigation actions towards the goal room that contains the target object. The dialog history is denoted as $< t_o, Q_1, A_1, Q_2, A_2, ..., Q_i, A_i >$, including the target object t_o, the questions Q and answers A till the turn i ($0 \leq i \leq k$, where k is the total number of Q-A turns from the beginning to the goal room). The agent, located in p_0, is trying to move closer to the goal room by inferring from the dialog history that happened before.

Distributed Actor-Learner Navigation Learning Framework. To train models for the various language grounded navigation tasks like VLN and NDH, we develop a distributed actor-learner learning infrastructure[2]. The framework design is inspired by IMPALA [8] and uses its off-policy correction method called V-trace to efficiently scale reinforcement learning methods to thousands of machines. The framework additionally supports a variety of supervision strategies important for navigation tasks such as teacher-forcing [9], student-forcing [9] and mixed supervision [19]. The framework is built using TensorFlow [1] and supports ML accelerators (GPU, TPU).

3 Language-Grounded Multitask Navigation

3.1 Generalized Navigation Model

We use the reinforced cross-modal matching (RCM) model [21] as the backbone and present a generalized navigation model in Figure 1 for multitask reinforcement learning, which we refer as Multitask-RCM (MT-RCM). To make the MT-RCM model generalizable and seamlessly transfer across tasks, we share all the learnable parameters for both VLN and NDH, including the joint word embedding layer, the language encoder, the trajectory encoder, the cross-modal attention module (CM-ATT), and the action predictor.

[2] The identity is not disclosed to respect the anonymity of the submission.

Table 1: Average agent progress towards goal location when trained using different rewards and mixed supervision strategy. t_o is the target object. Q_i and A_i are the question asked by the navigator and the answering response from the Oracle at turn i.

	Model	**Inputs**				**Goal Progress (m)**	
		t_0	A_i	Q_i	$A_{1:i-1}; Q_{1:i-1}$	Val Seen	Val Unseen
Baselines	Shortest-Path Agent					9.52	9.58
	Random Agent					0.42	1.09
	Seq-2-Seq [19]	✓				5.71	**1.29**
		✓	✓			6.04	2.05
		✓	✓	✓		6.16	1.83
		✓	✓	✓	✓	5.92	2.10
Ours	RCM (distance to goal location)	✓				4.18	0.42
		✓	✓			4.96	2.34
		✓	✓	✓		4.60	2.25
		✓	✓	✓	✓	5.02	2.58
	RCM (distance to goal room)	✓				**6.97**	1.25
		✓	✓			**6.92**	**2.69**
		✓	✓	✓		**6.47**	**2.69**
		✓	✓	✓	✓	**6.49**	**2.64**

3.2 Interleaved Multitask Data Sampling

To avoid overfitting to individual tasks, we adopt an interleaved multitask data sampling strategy to train the model. Particularly, each data sample within a mini-batch can be any data sample from either task, so that the VLN instruction-trajectory pairs and the NDH dialog-trajectory pairs are interleaved in a mini-batch though they may have different learning objectives.

3.3 Reward Shaping

Distance to Goal Location. Following prior art [20, 21], we first implement a discounted cumulative reward function R for the VLN and NDH tasks:

$$R(s_t, a_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'}), \text{ where } r(s_{t'}, a_{t'}) = \begin{cases} d(s_{t'}, v_{tar}) - d(s_{t'+1}, v_{tar}) & \text{if } t' < T \\ \mathbb{1}[d(s_T, v_{tar}) \leq d_{th}] & \text{if } t' = T \end{cases} \tag{1}$$

where γ is the discounted factor, $d(s_{t'}, v_{tar})$ is the distance between s_t and the target location v_{tar}, and d_{th} is the maximum distance from v_{tar} that the agent is allowed to terminate for success.

Distance to Goal Room. Different from VLN, NDH is essentially room navigation instead of point navigation because the agent is expected to reach a room that contains the target object. Suppose the goal room is occupied by a set of nodes $\{v_i\}_1^N$, we replace the distance function $d(s_t, v_{tar})$ in Eq. 1 with the minimum distance to the goal room $d_{room}(s_t, \{v_i\}_1^N)$ for NDH:

$$d_{room}(s_t, \{v_i\}_1^N) = \min_{1 \leq i \leq N} d(s_t, v_i) \tag{2}$$

4 Experiments

Implementation Details. The models were trained using a mixed learning strategy wherein some episodes in the batch used behavioral cloning [4] while the rest of the episodes used REINFORCE policy-gradient updates [22]. The language encoder uses a bidirectional-LSTM [17]. Similar to benchmark models [9, 21, 12, 11], the visual encoder at each time step t encodes a 360-degree panoramic view of the current location discretized into k view angles ($k = 36$, 3 elevations by 12 headings at 30-degree intervals). The action space, which is common to VLN and NDH tasks, consists of feasible navigable directions from the current location. The vocabulary used for joint-training is the union of the two tasks' vocabularies.

Table 2: Comparison of agent performance when trained separately *vs.* jointly on VLN and NDH tasks. All the reported results are averages of 3 independent runs.

		NDH Evaluation					**VLN Evaluation**				
Fold	**Task(s) Trained**	**Inputs for NDH**				**Goal Progress**	**PL**	**NE**	**SR**	**SPL**	**CLS**
		t_0	A_i	Q_i	$A_{1:i-1}; Q_{1:i-1}$	↑		↓	↑	↑	↑
Val Seen	NDH only	✓				**6.97**					
		✓	✓			**6.92**					
		✓	✓	✓		**6.47**					
		✓	✓	✓	✓	**6.49**					
	VLN only						10.75	5.09	52.39	48.86	63.91
	NDH+VLN	✓				3.00	11.73	4.87	54.56	52.00	65.64
		✓	✓			5.92	11.12	4.62	54.89	**52.62**	66.05
		✓	✓	✓		5.43	10.94	**4.59**	54.23	52.06	66.93
		✓	✓	✓	✓	5.28	10.63	5.09	**56.42**	49.67	**68.28**
Val Unseen	NDH only	✓				1.25					
		✓	✓			2.69					
		✓	✓	✓		2.69					
		✓	✓	✓	✓	2.64					
	VLN only						10.60	6.10	42.93	38.88	54.86
	NDH+VLN	✓				**1.69**	13.12	5.84	42.75	38.71	53.09
		✓	✓			**4.01**	11.06	5.88	42.98	40.62	54.30
		✓	✓	✓		**3.75**	11.08	5.70	44.50	39.67	54.95
		✓	✓	✓	✓	**4.36**	10.23	**5.31**	**46.20**	**44.19**	**54.99**

Reward Shaping for NDH task. As discussed in Sec. 3.3, we first performed studies to shape the reward for the NDH task. We replicate the setup used in Thomason et al. [19] for studies—goal progress (in meters) is used as the main evaluation metric, mixed supervision path is used for episodes with behavioral cloning and ablation studies are performed by restricting the access to dialog history by the navigation agents. The results in Table 1 indicate that our proposed reward that incentivizes the agent to get closer to the goal room outperforms the goal-oriented reward.

Multitask learning for VLN and NDH tasks. Table 2 shows the results of jointly training MT-RCM model on VLN and NDH tasks. Firstly, the MT-RCM model jointly trained on VLN and NDH indeed learns a more generalized navigation policy, which outperforms its single-task baseline by a large margin on the previous unseen environments of both tasks. Moreover, the gap between the agent's performance on previously seen and unseen environments is significantly reduced. For instance, when the full history of the dialog is provided to the agent, the performance gap on seen and unseen environments in the NDH task reduces from 2.84m to 0.92m. Similar behavior on VLN is observed as well. Secondly, we see a consistent and gradual increase in the success rate of the agent on the VLN task as it is trained on paths with more dialog history from the NDH task. This shows that the agent benefits from more complete information about the path, which in turn implies the importance given by the agent to the language instructions in the task. Thirdly, we note here that all the results on NDH validation unseen dataset beat the state-of-the-art results [19] by a large margin. When provided access to full dialog history, MT-RCM outperforms the state-of-the-art goal progress of 2.10m by more than 100%. At the same time, MT-RCM outperforms the equivalent RCM baseline [21] of 40.6% success rate by more than 13% (relative measure) on R2R validation unseen dataset. Finally, we conducted an experiment to study the importance of parameter sharing during multitask learning, which is illustrated in Table 3 in Appendix.

5 Conclusion

Our generalized language-grounded multitask model outperforms the equivalent baseline for the VLN task and establishes a new state of the art for the NDH task. The interleaved multitask data sampling strategy prevents overfitting to a single task and obtains simultaneous improvement on both the tasks. We believe that the multitask learning framework can further be extended by adapting MT-RCM on other language-grounded navigation datasets like Touchdown [5] and TalkTheWalk [7].

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016. URL `https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf`.

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv*, 2018. `arXiv:1807.06757 [cs.AI]`.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[4] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*, pages 103–129, Oxford, UK, UK, 1999. Oxford University. ISBN 0-19-853867-7. URL `http://dl.acm.org/citation.cfm?id=647636.733043`.

[5] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[6] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063, 2018.

[7] Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

[8] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1407–1416, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/espeholt18a.html`.

[9] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[10] Sachithra Hemachandra, Felix Duvallet, Thomas M Howard, Nicholas Roy, Anthony Stentz, and Matthew R Walter. Learning models for following natural language directions in unknown environments. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5608–5615. IEEE, 2015.

[11] Haoshuo Huang, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. Multi-modal discriminative model for vision-and-language navigation. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 40–49, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1605. URL `https://www.aclweb.org/anthology/W19-1605`.

[12] Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhães, Jason Baldridge, and Eugene Ie. Transferable representation learning in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision178(ICCV)*, 2019.

[13] Vihan Jain, Gabriel Magalhães, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*, 2019.

[14] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[15] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[16] Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. Learning to navigate in cities without a map. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2419–2430. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7509-learning-to-navigate-in-cities-without-a-map.pdf`.

[17] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45:2673–2681, 1997.

[18] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1268.

[19] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*, 2019.

[20] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–53, 2018.

[21] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[22] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL `https://doi.org/10.1007/BF00992696`.

[23] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

A Appendix

One of the main advantages of multitask learning is that under-represented tokens in each of the individual tasks get a significant boost in the number of training samples. Figure 2 illustrates that tokens with less than 40 occurrences end up with sometimes more than 300 occurrences during joint-training.

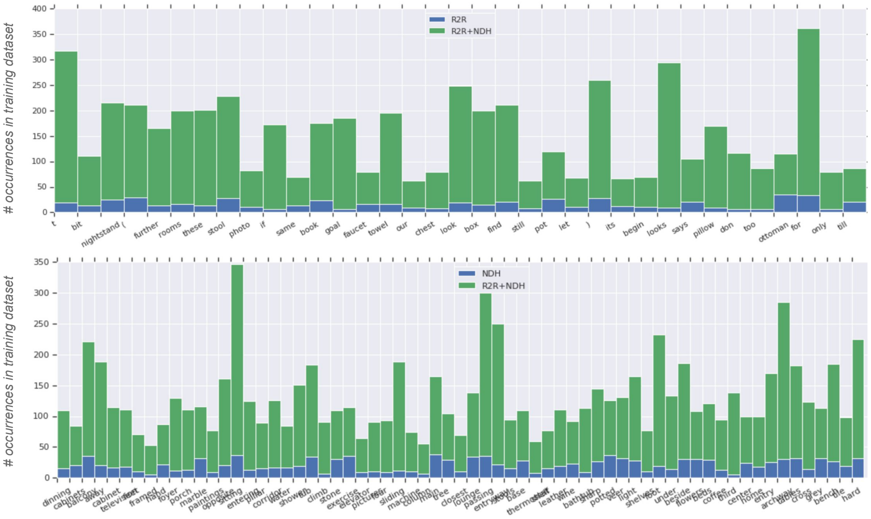

Figure 2: Selected tokens from the vocabulary for VLN (top) and NDH (bottom) tasks which gained more than 40 additional occurrences in the training dataset due to joint-training.

Table 3: Comparison of agent performance when language instructions are encoded by separate *vs.* shared encoder for VLN and NDH tasks. All the reported results are averages of 3 independent runs.

		NDH Evaluation					**VLN Evaluation**				
Fold	**Language Encoder**	**Inputs for NDH**				**Goal Progress**	**PL**	**NE**	**SR**	**SPL**	**CLS**
		t_0	A_i	Q_i	$A_{1:i-1}; Q_{1:i-1}$	↑		↓	↑	↑	↑
Val Seen	Shared	✓				**3.00**	11.73	4.87	54.56	52.00	65.64
		✓	✓			**5.92**	11.12	**4.62**	54.89	**52.62**	66.05
		✓	✓	✓		**5.43**	10.94	4.59	54.23	52.06	66.93
		✓	✓	✓	✓	**5.28**	10.63	5.09	**56.42**	49.67	**68.28**
	Separate	✓				2.85	11.43	4.81	54.66	51.11	65.37
		✓	✓			4.90	11.92	4.92	53.64	49.79	61.49
		✓	✓	✓		5.07	11.34	4.76	55.34	51.59	65.52
		✓	✓	✓	✓	5.17	11.26	5.02	52.38	48.80	64.19
Val Unseen	Shared	✓				1.69	13.12	5.84	42.75	38.71	53.09
		✓	✓			**4.01**	11.06	5.88	42.98	40.62	54.30
		✓	✓	✓		**3.75**	11.08	5.70	44.50	39.67	54.95
		✓	✓	✓	✓	**4.36**	10.23	**5.31**	**46.20**	**44.19**	**54.99**
	Separate	✓				**1.79**	11.85	6.01	42.43	38.19	54.01
		✓	✓			3.66	12.59	5.97	43.45	38.62	53.49
		✓	✓	✓		3.51	12.23	5.89	44.40	39.54	54.55
		✓	✓	✓	✓	4.07	11.72	6.04	43.64	39.49	54.57