

Multistage Variational Quantum Learning for Compositional Concept Generalization

Mina Abbaszadeh¹, Matilda Moore¹, Martha Lewis², Mehrnoosh Sadrzadeh¹

1. University College London 2. University of Amsterdam

Humans can understand novel situations by recombining previously learned concepts, a capability known as *compositional generalization*. For example, a person who has learned that cars crossing roads are dangerous will still recognize the danger posed by a yellow car, even if they have never encountered a yellow car before. Achieving this capability in AI remains an open challenge, particularly for learning relational concepts [1]–[4]. In principle, AI models based on compositional representations of meaning should exhibit stronger compositional generalization. One such framework is the Distributional Compositional Categorical model (DisCoCat) [5], which represents words with relational meanings as higher-order tensors, formed according to their semantic structure. Although theoretically well suited to compositional reasoning, DisCoCat has shown limited empirical success due to the computational cost of learning higher-order tensors on classical computers [6]. DisCoCat was inspired by Categorical Quantum Mechanics [7] and both frameworks are based on the theory of Compact Closed Categories. This enables us to represent DisCoCat constructions as Variational Quantum Circuits (VQCs), allowing for efficient training using hybrid quantum-classical optimization, leading to efficient execution on quantum hardware. VQCs avoid learning the tensors directly as multiple arrays of data, leading to a substantial improvement in scalability and practicality of DisCoCat [8], [9]. Despite these advances, DisCoCat has not been extended to compositional concept generalisation, and its extensions to multimodal learning remain limited, e.g. see here [10] for an extension to audio-text representations but only for adjectives noun phrases. Meanwhile, vision-language models such as CLIP [11] achieve strong performance in image-text alignment, but struggle with compositional generalization, particularly for relational reasoning [12], [13]. To address these limitations, we propose a variational quantum learning framework for DisCoCat, using a multistaged setting which first learns object representations from single object image-caption pairs, then freezes these and learns relational concepts.

The proposed model improves on previous work [14], with a single stage architecture that learnt the objects and relations simultaneously. Apart from its novel methodological contribution, this paper also: (a) provides an experimental study demonstrating improved relational generalization on the standard CLEVR benchmark [15], using the evaluation protocol of [12], [13] consisting of a difficult out of distribution testset, (b) consistently outperforms the classical CLIP baselines while using a fraction of parameters, and (c) presents a validation on noisy simulators and current quantum hardware.

DisCoCat to Variational Quantum Circuits. DisCoCat encodes conceptual relations by mapping the pregroup semantics of text to finite-dimensional Hilbert spaces via a homomorphic mapping. A pregroup is a partially ordered monoid where each element has a left and a right adjoint $(P, \cdot, 1, \leq, (-)^l, (-)^r)$. Composition is enabled via the monoid multiplication. Whenever an element and its adjoint compose in the correct order, a cancellation happens, that is we have $p \cdot p^r \leq 1$ and $p^l \cdot p \leq 1$, for 1 the unit of our monoid multiplication. The pregroup semantics of an atomic entity is a generator of the pregroup, that of a relation, on the other hand, consists of multiplied adjoint types. For instance, given a pregroup generated over the set $\{n, s\}$, the semantics of an atomic entity such as “cube” is n , whereas that of a relation such as “left of” or “right of” is $n^r \cdot s \cdot n^l$. When a relation composes with two entities, the following cancellation happens $n \cdot (n^r \cdot s \cdot n^l) \cdot n \leq s$. A pregroup-Hilbert Space mapping $\mathcal{F}$ sends atomic elements to finite dimensional Hilbert Spaces $\mathcal{F}(n) = V, \mathcal{F}(s) = W$, and elements with multiplied and adjoints to tensors of spaces with their duals $\mathcal{F}(n^r \cdot s \cdot n^l) = V^* \otimes W \otimes V^*$. As a result, composition becomes tensor product and cancellation tensor contraction. Due to space limitations we will not provide more details and refer the reader to existing work [5], [14], [16]. The Hilbert spaces that are assigned to entities and relations are easily transpiled to variational quantum circuits (VQCs) using an ansatz. We use Sim4 [17]. A few other options such Sim ansatz and IQP were experimented with, but did not perform as well.

Multistage Training. We encode the images as VQC’s using different methodologies, either by one-hot (OHE) or multi-hot (MHE) encoding, using five qubits per noun, or by frozen CLIP embeddings encoded as quantum states. For the latter, we consider both **angle encoding**, where PCA-reduced CLIP embeddings are mapped to

Models	Method	Train (avg)	Valid (avg)	Test (avg)	Best test	Trainable Parameters
OHE		92.20%	95.00%	93.34%	100%	168
CLIP	Angle Enc.	78.66%	75.00%	66.70%	70.31%	312
	Amplitude Enc.	87.37%	79.71%	80.97%	100%	312
(Classical Baseline) CLIP	CTP	95.63%	90.45%	91.00%	95.18%	63.4M

TABLE I: Results of single object training

Models	Method	Train (avg)	OOD Valid (avg)	OOD Test (avg)	Best Test	Parameters
MHE	Binary	99.00%	73.00%	85.88%	91.66%	279
Collage Encoding	Angle Enc.	82.20%	61.05%	72.32%	75.00%	426
	Ampl. Enc.	62.28%	40.00%	48.58%	50.00%	426
CLIP Multistage	Angle Enc.	76.60%	61.25%	55.33%	60.16%	480
	Ampl. Enc.	59.02%	45.00%	50.00%	50.00%	480
CLIP Non Multistage	Angle Enc.	55.26%	44.50%	53.09%	55.00%	792
	Ampl. Enc.	50.64%	62.00%	42.93%	48.33%	792
(Classical) CLIP	CTP	75.94%	11.14%	50.00%	50.00%	151.5M

TABLE II: Results of relational learning

rotation angles on 9 qubits, and **amplitude encoding**, where the full 512-dimensional embeddings are encoded as $|\psi\rangle = \sum_{i=1}^{512} x_i |i\rangle$. In order to compare with a classical baseline (**CTP**), we replace the quantum text encoder with a CLIP text encoder while keeping the CLIP image encoder frozen. In **Stage 1**, the model is trained on single object image-caption pairs using a supervised contrastive loss to learn grounded object representations shared across modalities. In **Stage 2**, the parameters learned for each shape in the previous stage are used as fixed rotations on the quantum circuits and only the relational parameters (*left*, *right*) are optimized using a binary image-caption alignment. We also introduce a **Collage** encoding, where relational image circuits are constructed by combining learned object circuits using entanglement.

Results and Analysis. All models are trained with the Adam optimizer. The results are averaged over five random seeds. They are summarized in Tables I and II. Overall, our framework consistently improves relational compositional generalization while requiring substantially fewer trainable parameters than both the non multistage quantum models and the classical CLIP baseline. For single object training, **OHE** achieves the highest quantum accuracy (93.34%), while amplitude encoding performs best among the CLIP-based quantum models (80.97%), using only 312 trainable parameters compared with 63.4M for CLIP. For **relational learning**, the classical CLIP baseline performs at chance level (50%), whereas all multistage quantum models achieve substantially higher accuracies using only a few hundred trainable parameters versus 151.5M for CLIP, with **MHE** obtaining the best overall result (85.88%) and the **Collage** encoding following suite at 72.32%. Among the CLIP-based quantum models, amplitude encoding was more effective for single objects, whereas angle encoding consistently yielded better relational generalization. We hypothesize that the nonlinear rotations of angle encoding produce more separable relational representations, contributing to an improved performance (see Fig. 2 in Appendix). Furthermore, multistage training consistently outperforms the corresponding non multistage models while reducing the number of trainable parameters (480 vs. 792), supporting the hypothesis that learning stable object representations before relational reasoning improves compositional generalization. Fine-tuning the CLIP image encoder did not improve the relational performance, consistent with previous findings [12].

Hardware Feasibility Evaluation. To evaluate the feasibility of executing the proposed framework on near-term quantum hardware, we perform hardware compatible similarity estimations using the destructive SWAP test using angle encoding and the CLIP Multistage model due to its lower dimensional image representations and generalisability; **OHE**, **MHE** and **Collage** are only feasible for an enumerable set of objects, such as those in

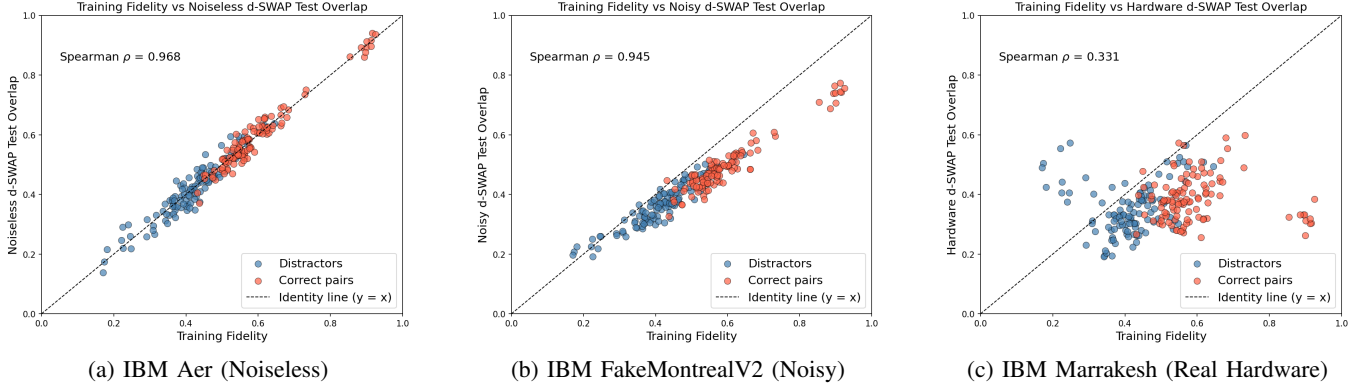


Fig. 1: Training fidelity versus destructive SWAP-test overlap across quantum emulators and hardware. Red points denote similar (correct) image-caption pairs, while blue points denote dissimilar (distractor) image-caption pairs.

CLEVR, whereas CLIP can be learnt on any image scraped from the web. The simulation experiments used 9-qubit image representations to match with the 512-dimensional CLIP embeddings ($2^9 = 512$ basis states), for the hardware experiments we had to reduce this to 4 qubits (see Table IV). During training, image-caption similarity is computed using the quantum state fidelity. After training, the optimized circuits are transpiled to the target backend, where the same fidelity is estimated using the destructive SWAP test.

Figure 1 shows excellent agreement between destructive SWAP estimates and exact fidelities for IBM Aer and IBM FakeMontrealV2. In contrast, agreement decreases on IBM Marrakesh, illustrating the impact of hardware noise and device constraints.

TABLE III: Comparison of destructive SWAP-test overlap estimates and classical quantum-state fidelities across 260 image-caption pairs. Mean overlaps, Spearman correlation, centroid distance and mean absolute error are reported.

Configuration	Dissimilar	Similar	ρ with classical	Centroid Distance	MAE
Relational Training Model (Classical)	0.413 ± 0.091	0.594 ± 0.113	—	0.133	—
IBM Aer (Noiseless), 1000 shots	0.414 ± 0.098	0.592 ± 0.118	0.968	0.132	0.023
IBM Aer (Noiseless), 100 shots	0.432 ± 0.129	0.602 ± 0.136	0.803	0.124	0.063
IBM FakeMontreal (Noisy), 1000 shots	0.364 ± 0.074	0.494 ± 0.092	0.945	0.096	0.076
IQM Aphrodite (Noisy), 1000 shots	0.175 ± 0.035	0.206 ± 0.038	0.585	0.022	0.313
IBM FakeMarrakesh (Noisy), 1000 shots	0.336 ± 0.070	0.471 ± 0.088	0.921	0.101	0.099
IBM Marrakesh (Real Hardware), 1000 shots	0.351 ± 0.085	0.385 ± 0.075	0.331	0.063	0.159

Consistent with these observations, Table III shows that increasing the number of shots improves agreement with the classical fidelities, while hardware noise largely preserves the separation between similar and dissimilar image-caption pairs. The IQM FakeAphrodite backend produces lower overlap values and weaker correlation than the IBM emulators. Inspection of the backend calibration data suggests that the stronger degradation is consistent with shorter median coherence times ($T_1 \approx 48.9$ us and $T_2 \approx 9.60$ us) together with higher gate (1-qubit gates: ≈ 0.001 , 2-qubit gates: ≈ 0.015) and readout error rates (≈ 0.05) compared to the IBM backends. Nevertheless, all backends retain a positive correlation (Spearman’s ρ) with the classical fidelities. Naturally, the real runs had the lowest correlations but as plot (c) of Fig. 1 shows, they still distinguish the similar and dissimilar pairs. In order to quantify this numerically, we computed the cluster centroid distances and a Mean Absolute Error (MAE), which although decreasing from noiseless simulations to real runs, stayed positive. These results show that the proposed framework can learn concept compositions and especially relational concepts and provides a good use case for current quantum hardware.

Future Work. We only trained the text component while keeping the visual representations fixed throughout, due to lacking a compositional theory for images. Future work focuses on constructing such a theory, translating it to VQCs, and experimenting on concept generalization.

REFERENCES

- [1] J. A. Fodor and Z. W. Pylyshyn, “Connectionism and cognitive architecture: A critical analysis,” *Cognition*, vol. 28, no. 1–2, pp. 3–71, 1988.
- [2] B. H. Partee, “Lexical semantics and compositionality,” *An invitation to cognitive science*, vol. 1, pp. 311–360, 1995.
- [3] P. Smolensky, “Tensor product variable binding and the representation of symbolic structures in connectionist systems,” *Artificial Intelligence*, vol. 46, no. 1–2, pp. 159–216, 1990.
- [4] E. Pavlick, “Semantic structure in deep learning,” *Annual Review of Linguistics*, vol. 8, no. 1, pp. 447–471, 2022.
- [5] B. Coecke, M. Sadrzadeh, and S. Clark, “Mathematical Foundations for a Compositional Distributional Model of Meaning,” *arXiv:1003.4394 [cs, math]*, Mar. 2010. [Online]. Available: <http://arxiv.org/abs/1003.4394>
- [6] M. Lewis, P. Smolensky, and M. Sadrzadeh, “Clip-binding: A systematic study of compositionality in vision–language models,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.
- [7] S. Abramsky and B. Coecke, “A categorical semantics of quantum protocols,” 2007. [Online]. Available: <https://arxiv.org/abs/quant-ph/0402130>
- [8] R. Lorenz, A. Pearson, K. Meichanetzidis, D. Kartsaklis, and B. Coecke, “QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer,” *arXiv:2102.12846 [quant-ph]*, 2021.
- [9] K. Meichanetzidis, A. Toumi, G. de Felice, and B. Coecke, “Grammar-aware sentence classification on quantum computers,” *Quantum Machine Intelligence*, vol. 5, no. 1, Feb. 2023.
- [10] S. Nazir and M. Sadrzadeh, “How does an adjective sound like? exploring audio phrase composition with textual embeddings,” in *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*. Association for Computational Linguistics, 2024, pp. 13–18.
- [11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763, iSSN: 2640-3498.
- [12] M. Lewis, N. Nayak, P. Yu, J. Merullo, Q. Yu, S. Bach, and E. Pavlick, “Does CLIP bind concepts? probing compositionality in large image models,” in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 1487–1500.
- [13] B. Pearson, M. Wray, and M. Lewis, “Diffusion models for improved compositional generalisation in vlms,” 2024.
- [14] H. Hawashin, M. Abbaszadeh, N. Joseph, B. Pearson, M. Lewis, and M. Sadrzadeh, “Compositional concept generalization with variational quantum circuits,” in *Proceedings of the Conference IEEE Quantum AI*. ACL, 2025. [Online]. Available: [arXiv:2509.09541](https://arxiv.org/abs/2509.09541)
- [15] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [16] R. Lorenz, K. Meichanetzidis, and B. Coecke, “Qnlp in practice: Running compositional models of meaning on a quantum computer,” *arXiv preprint arXiv:2102.04127*, 2021.
- [17] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, “Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms,” *arXiv preprint arXiv:1905.10876*, May 2019.

APPENDIX

TABLE IV: Performance of the reduced 4-qubit angle-encoding model used for hardware experiments.

Models	Method	Train	Valid	OOD-Test	Best Test	Parameters
Single-object training	Angle Enc.	94.19%	95%	93.34%	100.00%	132
Relational training	Angle Enc.	78.91%	77.46%	72.69%	73.84%	210

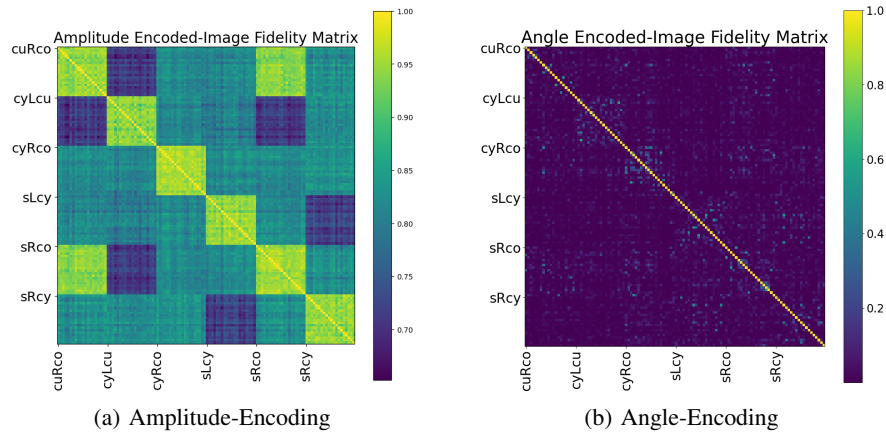


Fig. 2: Fidelity matrices for amplitude and angle encoding. cu=cube, cy=cylinder, co=cone, s=sphere, L/R=left/right.