

QuCLIP: A Language Informed Quantum Vision-Language Model

T. G. Limbäck-Stokin¹, G. Pinheiro², K.I. Lo¹, M. Sadrzadeh¹

¹University College London ²University of Edinburgh

Introduction Vision-language understanding is a key challenge in AI, with applications to image-text alignment and retrieval [17]. Models such as OpenAI’s CLIP [24] have shown that large-scale joint embeddings can effectively connect visual and textual data. However, they primarily rely on Transformer architectures with dense attention, which can overlook linguistic structure and consequently misunderstand the underlying meaning of the text [14]. Recent evaluations of CLIP-like models (e.g., ARO, SugarCrepe, Winoground, and SVO-probes [31, 11, 28, 10]) demonstrate that these models frequently ignore word order and predicate-argument structure, thus struggle with identifying hierarchical relations such as verbs, acting like bag-of-words models. Addressing these challenges purely *classically* requires pre-training with hard negatives, which typically relies on datasets exceeding 400 million image-text pairs [24] and models scaling beyond 10 billion parameters [19].

In contrast, Quantum Machine Learning (QML) can take structure into account by mathematically embedding data symmetries and relational constraints into the gate topology of Variational Quantum Circuits (VQCs). Examples include Quantum Convolutional Neural Networks (QCNNs) [6], Quantum Graph Neural Networks [9], and Equivariant Quantum Neural Networks [15], which preserve the symmetries in images, graphs, and feature orderings. While highly effective at capturing symmetries in molecular and physical systems, these models were not explicitly designed to handle the hierarchical structures inherent to natural language.

To bridge this gap, we introduce QuCLIP, the first quantum vision-language model that encodes linguistic structure using a specialized Combinatory Categorical Grammar (CCG) ansatz. CCG formalizes the hierarchical linguistic relationships by mapping syntax to semantics via function application [27]. Previous work translates functions into tensors and represents function application by tensor contraction [20]. The novel contributions of this paper are two-fold. On the theoretical front, (i) we develop a *quantum circuit* semantics for CCG, where the tensors become entangled spaces and contractions Bell tests over them, (ii) by restricting entanglement to adjacent physical positions for text, using unitaries instead of Bell tests and quantum feature maps for images, we design a hardware-aware ansatz that reduces shot count and noise. On the experimental side, we evaluate the setting on three compositional vision-language benchmarks (SVO-Swap, ARO-Att, ARO-Rel) via classical simulations, then test via emulation of quantum hardware and benchmark the results in the presence of noise.

The QuCLIP Architecture CCG consists of a basic and an advanced set of types and inference rules. The basic CCG has the set of atomic types $\{N, NP, S\}$ representing nouns, noun phrases, and sentences, and two function types: forward and backward application $A \backslash B$ and A/B . The formal definitions of these rules 1 and an example derivation for the sentence “*Alice likes Bob*” are presented on the right 2.

A *circuit semantics* for CCG is developed via a homomorphic **CCG-VQC** mapping n between the CCG derivations and VQCs, by induction over the CCG derivation trees. This map assigns to each atomic type A a number of qubits $n(A)$. The number of qubits assigned to function types X/Y or $X \backslash Y$ is defined recursively as follows $n(X/Y) = n(X \backslash Y) = n(X) + n(Y)$. The circuit semantics of a word with type T is an $n(T)$ -qubit quantum state $|\psi\rangle$, prepared by a variational quantum circuit $U_{n(T)}(\Theta)$ that acts on an initial $|0\rangle^{\otimes n}$ state where $\mathcal{H} = \mathbb{C}^2$, see equation 1 below:

$$\mathcal{G}(w) = |\psi_w\rangle = U(\Theta_w)|0\rangle^{\otimes q_w} \in \mathcal{H}^{\otimes (\sum_{j=1}^k |t_j|)} \quad (1)$$

The circuit semantics of a sentence consisting of a sequence of n words $w_1 w_2 \dots w_n$ is $\mathcal{G}(w_1 \dots w_n)$, obtained by preparing the tensor product of the individual word states and applying Bell measurements to contract the shared grammatical types. Let $q_i = \sum_{j=1}^k |t_j|$ be the number of qubits assigned to the word w_i . Suppose the CCG derivation tree dictates the set of function applications S , where each application $k \in S$ represents a shared type t_k between a subsystem A_k and a subsystem B_k . In this case, the circuit semantics of a sentence is defined in equation 2:

$$\mathcal{G}(w_1 \dots w_n) = \left(\bigotimes_{k \in S} \langle \Phi |^{\otimes |t_k|}_{A_k, B_k} \right) \left(\bigotimes_{i=1}^n U(\Theta_i) |0\rangle^{\otimes q_i} \right) \quad (2)$$

$$\frac{X/Y \quad Y}{X} < \quad \frac{Y \quad X \backslash Y}{X} >$$

(1) Forward and backward application.

$$\begin{array}{ccc} \text{Alice} & \text{likes} & \text{Bob} \\ \hline \overline{NP} & \overline{(S \backslash NP)/NP} & \overline{NP} \\ \hline & S \backslash NP & < \\ \hline & S & > \end{array}$$

(2) CCG derivation of transitive sentence.

where $|t_k|$ is the number of qubits corresponding to the atomic type t_k , and $\langle\Phi| = \frac{1}{\sqrt{2}}(\langle 00| + \langle 11|)$ is the Bell test applied to the matching pairs of wires.

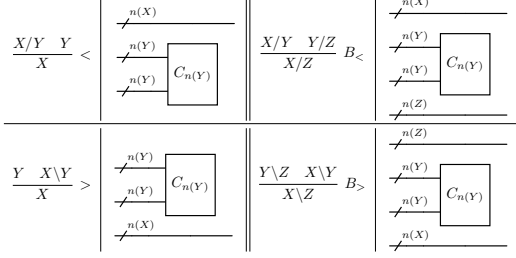


Figure 1: Mapping the CCG rules to VQCs. The C gate performs Bell measurements mirroring predicate-application 2.

A Hardware-Aware Ansatz Learning complex data structures requires intricate arrangements of entangling gates and this often clashes with the limited physical connectivity of near-term quantum hardware [3]. Mapping hierarchical relational structures into VQCs requires highly entangled tensor spaces, resulting in deep circuits that are susceptible to hardware noise and decoherence. [7]. To mitigate this, QuCLIP employs a set of hardware efficient ansatz schemas [12]. For the textual data, we utilize a slightly modified variant of the Sim14 [26] ansatz. It is defined sequentially as: (i) a layer of Hadamard gates H , (ii) a layer of general rotation gates R , (iii) a ring of CNOTs, (iv) a second rotation layer, and (v) a reverse ladder of CNOTs 2. Application rules are executed via Bell measurements followed by post selection [21], requiring all measured qubits to be in the $|0\rangle$ state [5]. Within each layer, a qubit undergoes a sequential three-parameter rotation block, granting the model full $SU(2)$ expressivity. Following these rotations, entanglement is generated via a periodic ring of CNOT gates coupling nearest-neighbours ($i \rightarrow (i + 1) \pmod{N}$). By restricting entangling operations strictly to adjacent physical positions, excluding the variable strides present in popular choices like the strongly entangling ansatz [25], this topology minimizes the need for costly qubit routing algorithms and avoids high-depth Swap gate overheads. The post selection process imposed by the Bell measurements causes the majority of circuit executions to be discarded, since the probability of a state $|0\rangle^{\otimes n_b}$ is only $1/2^{n_b}$, with n_b as the number of measured qubits, this results in an exponential growth in the total number of shots. We mitigate this by using the map-state duality and replace Bell tests by the direct application of the word unitaries [2]. Despite increasing the depth, this approach is necessary to reduce the exponential number of shots required per circuit.

For images, instead of the standard Amplitude Encoding (AE) [22], we use quantum feature maps. This is because the use of AE with the 512-dimensional CLIP embeddings is not viable with the available quantum hardware as the large depth results in a full decoherence of the state. The use of quantum feature maps significantly reduces the depth footprint typically associated with standard amplitude encoding with minimal changes in performance 3. We implemented two quantum feature map approaches [23]: firstly we implemented PCA reduced vectors directly in the ansatz, secondly, we trained a Multilayer Perceptron (MLP) to project embeddings to a set of features and used the resulting embeddings on the ansatz. In order to maintain the same circuit size across all methods, we implemented feature re-uploading on 9 qubits.

Objective Function Unlike classical models like CLIP that minimize the cosine distance between classical embeddings [30], QuCLIP aligns quantum state vectors. To effectively train this model, we introduce a Fubini-Study metric-inspired loss [8, 4] 3. We find this formulation outperforms classical cosine similarity and quantum fidelity by preserving gradients during contrastive learning, for more details see [18].

$$\mathcal{L} = - \sum_{i=1}^N \log \frac{\exp(s(I_i, T_i)/t)}{\sum_{j=1}^N \exp(s(I_i, T_j)/t)}, \quad (3)$$

$$s(I, T) = \arcsin(|\langle \psi_{\text{txt}} | \psi_{\text{img}} \rangle|)$$

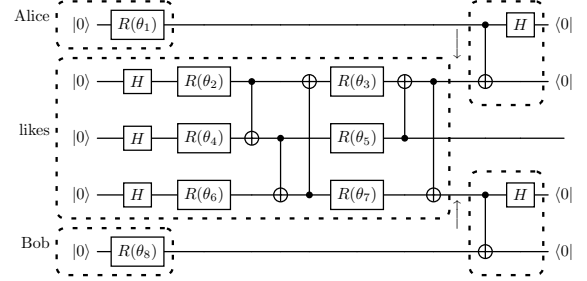


Figure 2: The quantum circuit for the sentence ‘‘Alice likes Bob’’ with $n(N) = n(S) = 1$.

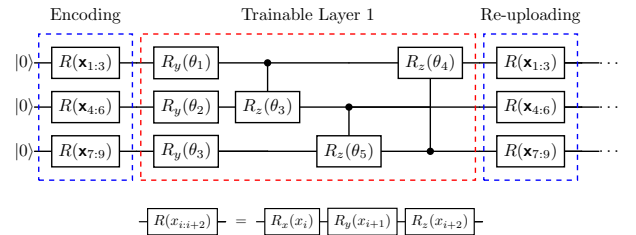


Figure 3: Quantum feature map architecture. Top: 3-qubit ansatz with alternating feature re-uploading and parametrised VQC layers. Bottom: The single-qubit three feature embedding block $R(\phi_i)$.

Experiments & Results We evaluated the model on two benchmarks: SVO-Swap and ARO, with ARO-Att and ARO-Rel subsets. SVO-Swap was derived from the bigger SVO-Probes dataset. For SVO-Probes, we used a 60/20/20 train/-validation/test split, excluding the entries selected for the SVO-Swap subset. For the ARO datasets, we use a 70/15/15 split for both the attribution and relation variants, then combined the training and validation splits while keeping the tests separate. Training was performed on an RTX 3090 Ti GPU with 24 Gb VRAM and 128 Gb standard RAM, we used the Adam optimiser [13] with a learning rate of $1e^{-3}$ for VQC parameters and $1e^{-5}$ for classical parameters present in the MLP feature map. Each text model uses the above ansatz with one qubit per type, except for the output which is assigned 9 qubits, and 2 layers. For the image side, we used the quantum feature map with 9 qubits and applied the interleaved feature uploading along with a variational layer 4 times. The circuits are composed via the compute-uncompute method, where the complex adjoint of the image circuit is appended to the text circuit and the probability of the 0 state is measured.

Setting	Model	Shots	ARO-Att		ARO-Rel		SVO-Swap		Overall
			Acc (%)	ρ	Acc (%)	ρ	Acc (%)	ρ	
Classical Simulation	PCA	–	47.85	1.000	49.80	1.000	69.47	1.000	0.000
	MLP	–	54.12	1.000	53.26	1.000	77.83	1.000	0.000
Noiseless Emulation	PCA	1000	51.74	0.885	50.14	0.837	73.68	0.959	0.0296
	MLP	1000	51.12	0.819	50.58	0.895	83.16	0.964	0.0242
Noisy Emulation	PCA	100	48.47	0.232	47.80	0.207	67.37	0.891	0.0330
		1000	49.11	0.185	49.86	0.154	70.53	0.784	0.0328
	MLP	100	48.86	0.279	50.66	0.066	77.89	0.702	0.0483
		1000	51.37	0.264	51.68	0.055	81.05	0.719	0.0501

The results show a comparative performance on all three benchmarks across classical simulations, noiseless quantum simulations and noisy runs. For these latter, we used the **FakeMiami**, an emulator of the quantum computer IBM Miami to run the test sets. Moreover, we establish bounds for correlation and macro mean absolute error between the classical baseline and its corresponding emulation as $\rho = 1.000$ and $\mathcal{D}_{\text{macro}} = 0.000$, respectively. For the classical simulations, MLP achieves the highest accuracy, peaking at 54.12% on ARO-Att, 53.26% on ARO-Rel, and 77.83% on SVO-Swap. For the quantum runs, the noiseless MLP deployment still peaked on SVO-Swap at 83.16%, outperforming its classical counterpart and highlighting the structural advantage of the language-informed circuit topology. On the ARO benchmarks, noiseless PCA and MLP performed very similarly, reaching an accuracy between 50.14% and 51.74%. Under noisy conditions, the performance of both degraded. However, scaling the execution from 100 to 1000 shots consistently mitigated the degradation and recovered accuracy across nearly all evaluations: the accuracy of noisy MLP model on ARO-Att increased from 48.86% to 51.37%, on ARO-Rel from 50.66% to 51.68%, on SVO-Swap from 77.89% to 81.05%. The trend was similar for PCA, but at 1000 shots it stayed slightly below MLP. Tracking the correlation metric (Spearman’s ρ) between classical simulation outcomes and noisy/noiseless emulations revealed similar results: correlations were positive across all models and datasets, but varied slightly between benchmarks, where occasionally the MLP does better and PCA underperformed. the underlying margin gaps between correct and incorrect predictions change only slightly in the presence of noise, as captured by the bounded macro mean absolute error ($\mathcal{D}_{\text{macro}}$) in the final column.

Discussion & Future Work QuCLIP provides a use case for current quantum hardware and an application of QML to real-world, large-scale multimodal data, moving away from restricted datasets such as MNIST or Bars and Stripes [16, 1]. It enforces linguistic structure via a homomorphic **CCG-VQC** mapping. To ensure NISQ viability we utilizes map-state duality, hardware aware ansatzes, and quantum feature maps with data reuploading to reduce qubit count, depth, and post selection overheads. The results are comparable across classical simulations vs noiseless and noisy emulations.

Our hardware aware ansatz only outperformed CLIP on SVO-Swap. This is due to having to work with a small number of qubits and depth to keep the hardware costs and noise down. In the absence of these restrictions with an average number of 12 qubits and depth of 108, the model outperformed CLIP and its more recent variant OpenCLIP on all three benchmarks (see Table 1 of [18]). We are working with UK’s National Quantum Computing Centre and its SparQ Access Programme to access their elaborate hardware platforms to remedy this problem. QuCLIP lacks compositional representations for images. Ongoing work bridges this gap by integrating image parsers [29] to decompose images into hierarchical scene graphs, mapping the visual relations onto VQC topologies and opening pathways for early fusion mechanisms.

References

- [1] Marcello Benedetti, Delfina Garcia-Pintos, Alejandro Perdomo-Ortiz, Vicente Evangelista, Yudong Cao, and Roberto Alejandro-Plaza. A generative modeling approach for benchmarking quantum hardware. *Npj Quantum Information*, 5(1):45, 2019.
- [2] Ville Bergholm. A diagrammatic approach to map-state and channel-state dualities. *arXiv preprint arXiv:1509.08339*, 2015.
- [3] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, August 2021.
- [4] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1), March 2021.
- [5] Bob Coecke, Giovanni de Felice, Konstantinos Meichanetzidis, and Sean Wang. Foundations for compositional quantum natural language processing. *arXiv preprint arXiv:2012.01255*, 2020.
- [6] Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15:1273–1278, 2019.
- [7] Alexander Cowtan, Silas Dilkes, Ross Duncan, Alexandre Krajenbrink, Will Simmons, and Seyon Sivarajah. On the qubit routing problem. In *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*, pages 5:1–5:32, 2019.
- [8] Marwan Ait Haddou and Mohamed Bennai. Sculpting quantum landscapes: Fubini-study metric conditioning for geometry aware learning in parameterized quantum circuits, 2025.
- [9] Noha Hassan, Xavier Fernando, and Halim Yanikomeroglu. Quantum graph neural networks for double-sided reconfigurable intelligent surface optimization. *arXiv preprint arXiv:2604.10453*, 2026.
- [10] Lisa Anne Hendricks and Aida Nematzadeh. Probing image–language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online, Aug 2021. Association for Computational Linguistics.
- [11] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023.
- [12] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Jerry M. Brink, Markus Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.
- [13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [14] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, 2018.
- [15] Isabel Nha Minh Le, Oriel Kiss, Julian Schuhmacher, Ivano Tavernelli, and Francesco Tacchino. Symmetry-invariant quantum machine learning force fields. *New Journal of Physics*, 27:023015, 2023.
- [16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [18] Tilen Gaetano Limbäck-Stokin, Tanishka Birdavade, Kin Ian Lo, and Mehrnoosh Sadrzadeh. Meaning representations as variational quantum circuits. In *Proceedings of the 7th International Workshop on Designing Meaning Representations (DMR)*, pages 113–123. European Language Resources Association (ELRA), 2026.

- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [20] Kin Ian Lo, Hala Hawashin, Mina Abbaszadeh, Tilen Gaetano Limbäck-Stokin, Hadi Wazni, and Mehrnoosh Sadrzadeh. DisCoCLIP: A distributional compositional tensor network encoder for vision-language understanding. In Lea Frermann and Mark Stevenson, editors, *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 316–327, Suzhou, China, November 2025. Association for Computational Linguistics.
- [21] Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. Qnlp in practice: Running compositional models of meaning on a quantum computer, 2021.
- [22] Mikko Mottonen, Juha J Vartiainen, Ville Bergholm, and Martti M Salomaa. Transformation of quantum states using uniformly controlled rotations. *arXiv preprint quant-ph/0407010*, 2004.
- [23] Adrián Pérez-Salinas, Alba Cervera-Liarta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020.
- [24] 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. Learning transferable visual models from natural language supervision, 2021.
- [25] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Killoran. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.
- [26] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12), October 2019.
- [27] Mark Steedman and Jason Baldridge. *Combinatory Categorical Grammar*, chapter 5, pages 181–224. John Wiley & Sons, Ltd, 2011.
- [28] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022.
- [29] Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005.
- [30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.
- [31] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.